HARNESSING DATA INNOVATION FOR MIGRATION POLICY

A HANDBOOK FOR PRACTITIONERS

IOM UN MIGRATION | GLOBAL DATA INSTITUTE

This publication was made possible through support provided by the German Federal Ministry of the Interior and Community. The opinions expressed herein are those of the authors and do not necessarily reflect the views of IOM.

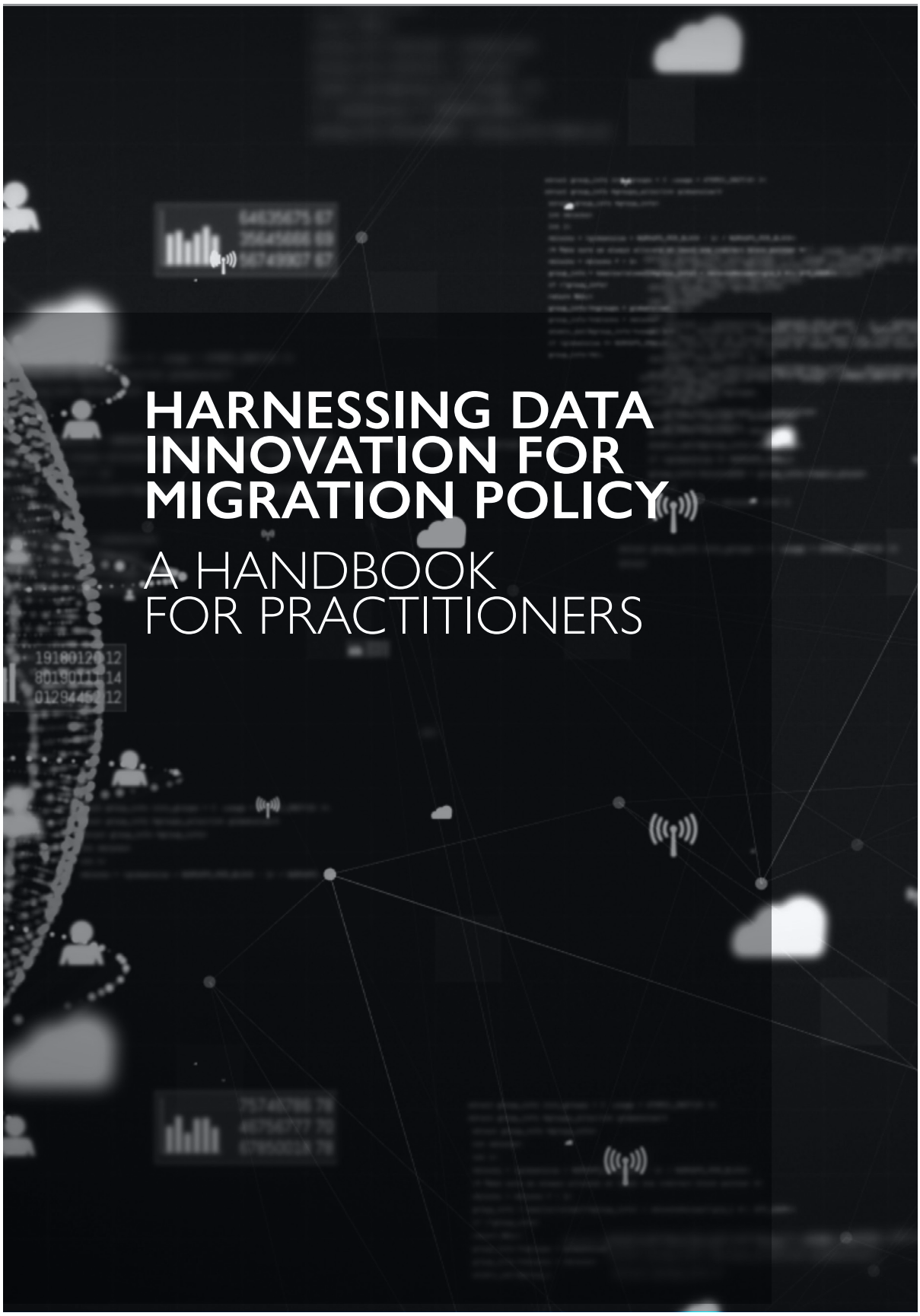
Publisher: International Organization for Migration
17 route des Morillons
P.O. Box 17
1211 Geneva 19
Switzerland
Tel.: +41 22 717 9111
Fax: +41 22 798 6150
Email: hq@iom.int
Internet: www.iom.int

Editors: Marzia Rango, Niklas Sievers and Frank Laczko
Support editing: Alina Menocal and Damien Jusselme
Art direction and data visualization: Roberta Aita
Language editor: Laarni Alfaro
Layout artists: Ramir Recinto and Mae Angeline Delgado

Required citation: International Organization for Migration (IOM), 2023. *Harnessing Data Innovation for Migration Policy: A Handbook for Practitioners*. IOM, Geneva.

ISBN 978-92-9268-444-0 (PDF)
ISBN 978-92-9268-445-7 (print)

PUB2022/091/EL



HARNESSING DATA INNOVATION FOR MIGRATION POLICY

A HANDBOOK FOR PRACTITIONERS

IOM UN MIGRATION | GLOBAL DATA INSTITUTE

BD4M BIG DATA FOR MIGRATION ALLIANCE

European Commission

KCMD

GOVLAB

FOREWORD

By the International Organization for Migration
Ugochi Daniels, IOM Deputy Director General for Operations

At a time of exponential data growth worldwide, with mobile phones, satellites and social media, along with new analytical methods, such as machine learning, these innovative data sources can augment traditional sources of evidence for migration policy and practice. The potential of these sources to fill migration data gaps remains largely untapped and unexplored, and raises questions in need of consideration.

This publication provides a tool for policymakers, officials, practitioners, academia and data holders to harness big data and innovative methods in a responsible and ethical way. The *Practitioners' Handbook* first sheds light on the solutions that big data and machine learning can bring to the analysis of migration. The *Handbook* shares methods to estimate flows and stocks, analyse the socioeconomic characteristics of migrants, and identify drivers. It also discusses forward-looking approaches to data analytics (forecast and foresight), and the assessment of public opinion on topics related to migration.

Furthermore, the *Handbook* highlights the challenges in using innovative, non-traditional data sources for migration-related trends. Chapters discuss how to deal with noise in big data, representativeness and bias. Large sets of data made possible by innovation in data mining raise questions about how to validate results against official sources of information. The *Handbook* provides good practice guidelines about data governance collaborations, which have to ensure the security and privacy of data from individuals, among other tasks.

IOM, through the Global Data Institute (GDI), seeks new ways to engage with different actors involved in the migration data journey. The Big Data for Migration (BD4M) Alliance, convened by GDI, the European Commission Joint Research Centre and The GovLab, is the first dedicated network of stakeholders seeking to facilitate responsible data innovation and build partnerships to improve the evidence base on migration and human mobility and its use for policymaking.

This *Handbook*, an endeavour of IOM and the BD4M Alliance, aims to accelerate the use of non-traditional data sources and innovative methods to support current and future migration-related programming and policy at the global, national and local levels. This resource provides tools and methods for harnessing data innovation for migration policy and practice.

By the European Union Joint Research Centre

Since the first decade of the last millennium, we have been witnessing a time when data production is omnipresent, and data storage is affordable and widely accessible. The proliferation of smartphones, connected devices and sensors has led to the collection of various forms of data on human behaviour, known as "digital traces". These specific conditions in data production and storage and processing technologies have enabled researchers to analyse the data sets for scientific purposes. One area where this "data revolution" has had the most significant impact is the social sciences, where the birth of computational social science, a new discipline that employs computational methods to study human behaviour using large-scale data, can be seen as its paradigmatic example.

The mainstreaming and use of these techniques in policymaking has a sure and evident potential, but it also presents significant challenges. Without any claim of exhaustivity, we may mention data access, privacy concerns and ethical considerations – and also the suitability and interpretability of analytical techniques, the design of data-driven policymaking processes, and mapping the kind of questions that can be tackled via the use of innovative data and analytical techniques.

A strand of work, animated by the European Commission as well as by other stakeholders, has been devoted to tackling these kind of issues: from the mapping of policy-relevant research questions, which can be addressed using innovative data and novel modelling techniques,[1] to specific analyses of data innovation in human mobility and demography,[2] as well as handbooks on mapping knowledge and proposing ideas and solutions for the application of computational social science to policymaking.[3]

Thanks to its specific role of knowledge broker, and as a place of exchange of knowledge between experts, practitioners and academics, the Big Data for Migration (BD4M) Alliance is particularly suited to contribute to this interesting field, and we believe that the present *Practitioners' Handbook* would greatly help in providing policymakers active in migration management with a toolbox able to tackle issues present in their daily work.

The BD4M Alliance exists to accelerate and nurture data collaboration. By fostering a community of experts and practitioners focused on harnessing new data sources for migration, the Alliance aims to facilitate knowledge exchange, capacity-building and the development of best practices in the field. The recently launched *Handbook*, in part created by the BD4M Alliance, serves as a valuable resource for professionals who are seeking to better understand and implement data collaboratives in the context of migration policy and research.

We believe the present *Practitioners' Handbook* to be a great and fundamental addition to the literature.

[1] Eleonora Bertoni, Matteo Fontana, Lorenzo Gabrielli, Serena Signorelli and Michele Vespe (eds.), *Mapping the Demand Side of Computational Social Science for Policy*, Publications Office of the European Union (2022). Available at https://publications.jrc.ec.europa.eu/repository/handle/JRC126781.

[2] Claudio Bosco, Sara Grubanov-Boskovic, Stefano M. Iacus, Umberto Minora, Francesco Sermi and Spyridon Spyratos, *Data Innovation in Demography, Migration and Human Mobility: Opportunities and Challenges of Non-traditional Data*, Publications Office of the European Union (2022). Available at https://op.europa.eu/en/publication-detail/-/publication/5282ef16-84a7-11ec-8c40-01aa75ed71a1.

[3] Eleonora Bertoni, Matteo Fontana, Lorenzo Gabrielli, Serena Signorelli and Michele Vespe (eds.), *Handbook of Computational Social Science for Policy*, Springer (2023). Available at https://link.springer.com/book/10.1007/978-3-031-16624-2.

The need for data collaboration

Stefaan G. Verhulst

In an increasingly interconnected world, migration has become a central and defining issue that demands informed and well-structured policies. How these policies are developed is of utmost importance as they directly impact the lives of millions of people, shaping societies and economies around the globe. To ensure their legitimacy, effectiveness and adaptability to rapidly changing circumstances, it is essential to leverage new data sources and utilize novel methodologies that can help unlock data in a systematic, sustainable and responsible manner.

In recent years, digital transformation has led to an unprecedented amount of data being generated every day. These new data sources, when combined responsibly with traditional ones, hold great potential for informing migration policies that can better address the complex and multifaceted challenges that surround human mobility. However, harnessing this potential requires the development and responsible implementation of innovative operational models and methodologies that can effectively tap into these vast troves of data.

As migration patterns continue to evolve, it is crucial to establish new partnership approaches that can foster access to and integration of diverse data sources. One such approach is the concept of data collaboratives, which bring together stakeholders from various sectors to exchange data, knowledge and expertise. These collaborations enable the pooling of resources and expertise that can significantly enhance our understanding of migration dynamics and inform evidence-based policymaking. The Big Data for Migration Alliance exists to generate and support these data collaboratives.

As we move forward, it is important that we continue to build on these initial efforts to develop a robust community of expertise and practice around the use of new data sources for migration. By encouraging the professionalization of data collaboratives, we can help ensure that migration policies are grounded in solid empirical evidence and tailored to the complex realities of today's world. The benefits of such an approach are clear: more informed and effective migration policies that can ultimately contribute to a more inclusive, equitable, diverse and prosperous global society.

ACKNOWLEDGEMENTS

This publication is the result of a highly collaborative effort involving several partners across sectors, countries, and disciplines participating in IOM Global Data Institute's exploratory research on harnessing innovative data sources, methods, and tools for migration policy and research.

The editors would like to sincerely thank all the contributing authors for making this publication possible, and all the reviewers for their constructive feedback on the draft chapters.

We want to extend special thanks to Marina Manke for her overall guidance and comments upon her appointment as the Chief of IOM's Global Migration Data Analysis Centre (GMDAC).

We are also grateful to the IOM Publications Unit, led by Valerie Hagger: Laarni Alfaro, Ramir Recinto and Mae Angeline Delgado. Further, we are indeed fortunate to have the backing of IOM GMDAC's Communications Team, led by Jorge Galindo, and assisted by Andi Armia Pratiwi and Roberta Aita. Finally, we would like to thank Frances Solinap, Tristan O'Shea and Paulina Kluczynska for their precious administrative support.

We are especially grateful to the German Federal Ministry of the Interior and Community for their financial support towards the completion of this report.

PARTNERS


IOM • OIM
DTM
THE UN MIGRATION AGENCY


euaa
EUROPEAN UNION
AGENCY FOR ASYLUM


European
Commission


FRA
EUROPEAN UNION AGENCY
FOR FUNDAMENTAL RIGHTS


gesis
Leibniz-Institut
für Sozialwissenschaften


Government
of Indonesia


GLOBAL
PULSE


GOVLAB


facebook


KCMD


JRC
EUROPEAN COMMISSION


MAX PLANCK
GESELLSCHAFT


positium


QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY


Telefónica


UNHCR
The UN Refugee Agency


United States®
Census
Bureau


UNIVERSITY OF
LIVERPOOL


IN SUPREMÆ DIGNITATIS
1343
UNIVERSITÀ DI PISA


UNIVERSITY OF
Southampton


Universiteit Utrecht


VUB
VRIJE
UNIVERSITEIT
BRUSSEL


WorldPop

CONTENTS

I HARNESSING DATA INNOVATION FOR MIGRATION POLICY: THEMATIC APPLICATIONS

HOW TO READ THIS HANDBOOK

This Guide is structured in a way to make the reader able to easily recognize the main findings of each chapter in the field of new data sources and policy purpose.

The icons will be placed at the beginning of each chapter.

CHAPTER NUMBER

6

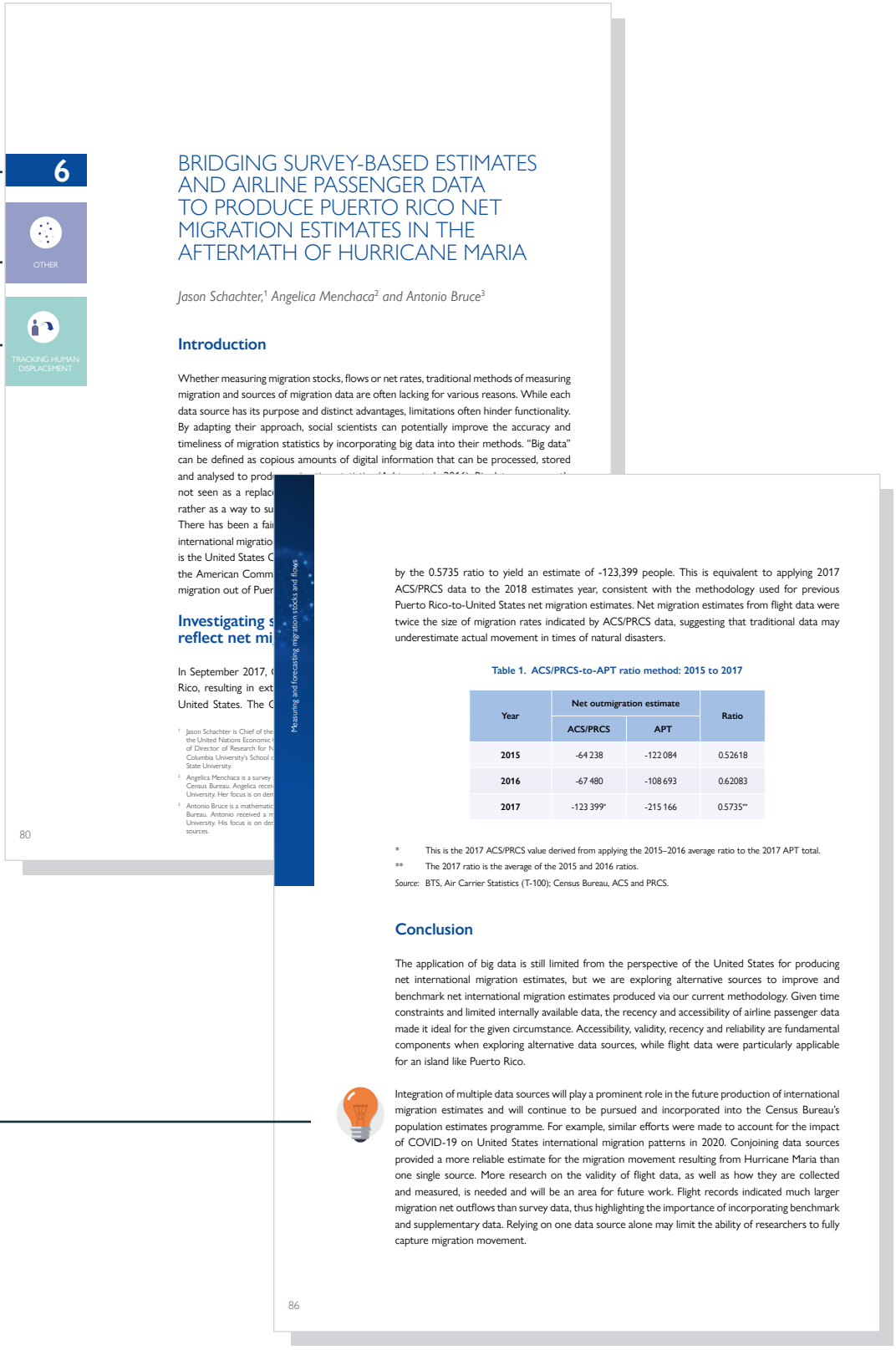
NEW DATA SOURCES

OTHER

PURPOSE

TRACKING HUMAN DISPLACEMENT

HIGHLIGHTS



BRIDGING SURVEY-BASED ESTIMATES AND AIRLINE PASSENGER DATA TO PRODUCE PUERTO RICO NET MIGRATION ESTIMATES IN THE AFTERMATH OF HURRICANE MARIA

Jason Schachter,[1] Angelica Menchaca[2] and Antonio Bruce[3]

Introduction

Whether measuring migration stocks, flows or net rates, traditional methods of measuring migration and sources of migration data are often lacking for various reasons. While each data source has its purpose and distinct advantages, limitations often hinder functionality. By adapting their approach, social scientists can potentially improve the accuracy and timeliness of migration statistics by incorporating big data into their methods. "Big data" can be defined as copious amounts of digital information that can be processed, stored and analysed to prod

not seen as a replac

rather as a way to su

There has been a fai

international migratio

is the United States C

the American Comm

migration out of Puer

Investigating s

reflect net mi

In September 2017, C

Rico, resulting in ext

United States. The C

[1] Jason Schachter is Chief of the

the United Nations Economic

of Director of Research for N

Columbia University's School

State University.

[2] Angelica Menchaca is a survey

Census Bureau. Angelica recei

University. Her focus is on der

[3] Antonio Bruce is a mathemati

Bureau. Antonio received a m

University. His focus is on der

sources.

80

Measuring and forecasting migration stocks and flows

by the 0.5735 ratio to yield an estimate of -123,399 people. This is equivalent to applying 2017 ACS/PRCS data to the 2018 estimates year, consistent with the methodology used for previous Puerto Rico-to-United States net migration estimates. Net migration estimates from flight data were twice the size of migration rates indicated by ACS/PRCS data, suggesting that traditional data may underestimate actual movement in times of natural disasters.

Table 1. ACS/PRCS-to-APT ratio method: 2015 to 2017

Year	Net outmigration estimate		Ratio
	ACS/PRCS	APT	
2015	-64 238	-122 084	0.52618
2016	-67 480	-108 693	0.62083
2017	-123 399*	-215 166	0.5735**

* This is the 2017 ACS/PRCS value derived from applying the 2015–2016 average ratio to the 2017 APT total.

** The 2017 ratio is the average of the 2015 and 2016 ratios.

Source: BTS, Air Carrier Statistics (T-100); Census Bureau, ACS and PRCS.

Conclusion

The application of big data is still limited from the perspective of the United States for producing net international migration estimates, but we are exploring alternative sources to improve and benchmark net international migration estimates produced via our current methodology. Given time constraints and limited internally available data, the recency and accessibility of airline passenger data made it ideal for the given circumstance. Accessibility, validity, recency and reliability are fundamental components when exploring alternative data sources, while flight data were particularly applicable for an island like Puerto Rico.

Integration of multiple data sources will play a prominent role in the future production of international migration estimates and will continue to be pursued and incorporated into the Census Bureau's population estimates programme. For example, similar efforts were made to account for the impact of COVID-19 on United States international migration patterns in 2020. Conjoining data sources provided a more reliable estimate for the migration movement resulting from Hurricane Maria than one single source. More research on the validity of flight data, as well as how they are collected and measured, is needed and will be an area for future work. Flight records indicated much larger migration net outflows than survey data, thus highlighting the importance of incorporating benchmark and supplementary data. Relying on one data source alone may limit the ability of researchers to fully capture migration movement.

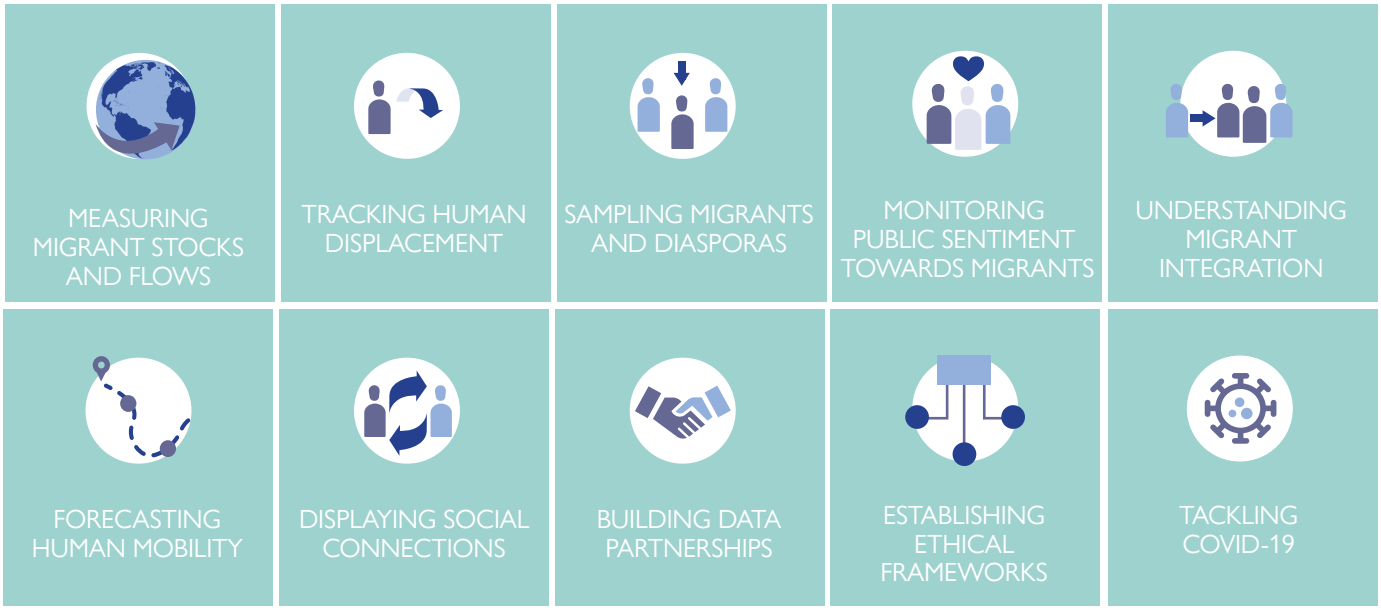
86

NEW DATA SOURCES

In this Handbook, the term "data innovation" will be used to refer to new data sources and innovative analytical methodologies offered by new technologies for the analysis of distinct policy-relevant migration issues. The data sources and methods can be organized broadly into the following groups:

PURPOSE

The new data sources and innovative analytics methodologies hold insights that are significant for a range of specific policy-relevant migration issues. These can be grouped into the following purposes for deploying new data sources and methods for migration policy:



HIGHLIGHTS

The highlight icon will be placed next to paragraphs containing info about the following topics:

DATA ACCESS	DATA PRIVACY
INFORMED CONSENT	GEOGRAPHICAL ACCURACY
NEAR REAL TIME	DATA SECURITY

INTRODUCTION

Marzia Rango and Niklas Sievers[1]

Why this handbook

Migration has risen as one of the most challenging issues confronting policymakers worldwide. The growing complexity of internal and cross-border human mobility has pointed out the need for accurate, timely and reliable information to develop migration policy, a need that is also reflected in the first objective of the Global Compact for Safe, Orderly and Regular Migration – "Collect and utilize accurate and disaggregated data as a basis for evidence-based policies" – and in the 2030 Agenda for Sustainable Development. However, traditional data sources such as surveys, censuses and administrative data are often not well suited to meet these needs.

National population censuses are the bedrock of migration statistics in most countries, but they are generally infrequent and (provided migration-related questions are included) mainly provide a static picture of the migration situation in a country at the time of data collection. Household surveys are detailed exercises, but they are very costly and affected by issues of representativeness of the migrant population. Administrative records, such as visa applications or work permits, hold significant potential for providing a dynamic picture of migration, but they are hardly used to produce migration statistics in most countries, due to quality and capacity issues. Only about 50 countries around the world have been collecting and reporting data on migration movements to the United Nations Statistics Division since 2010.[2] Data on migrants' profiles, including age, gender and socioeconomic profiles, are even patchier.

Meanwhile, rapid technological advancements in recent years have meant an exponential increase in the amount of data available globally. In 2018, this volume reached an estimated 33 zettabytes – equal to the compound storage of 33 billion 1-terabyte laptops. By 2021, this amount had doubled to 71 zettabytes, and it is expected to grow to a total of almost 180 zettabytes by 2025. New data sources such as social media, mobile phone data, and satellite imagery offer vast

[1] At the time of writing, Marzia Rango is Data Innovation and Capacity-Building Lead of IOM's Global Migration Data Analysis Centre (GMDAC). Niklas Sievers is Data Innovation Officer of IOM GMDAC.

[2] More information is available at www.migrationdataportal.org/themes/international-migrant-stocks (accessed 2 May 2022).

and diverse amounts of data relevant for migration research and governance. They can complement traditional data sources and help overcome some of the common challenges of traditional data-collection systems, such as relatively high costs of data collection (particularly for large-scale surveys), lack of timeliness, difficulties in measuring certain kinds of human mobility (particularly those of temporary and seasonal nature), and limited coverage of hard-to-reach populations.

Technological innovation is already changing the way we produce and use evidence on migration; the issue is how to do so more systematically, ethically and responsibly. The Big Data for Migration Alliance (BD4M) – convened in 2018 by IOM's Global Migration Data Analysis Centre, the European Commission's Knowledge Centre on Migration and Demography (KCMD), and the Governance Lab at New York University (The GovLab) – has implemented several activities to accelerate the ethical and responsible use of innovative data sources, tools, and methods to advance more humane and effective migration policies.[3]

Despite their evident potential and the fast-growing number of applications,[4] new data sources are far from being fully harnessed to tackle existing information gaps. Countries and organizations globally have not been able to implement these innovations at scale to design informed migration programmes and policies. This is due to a series of difficulties, ranging from limited technical capacities and skills, to a lack of adequate regulatory frameworks to safeguard data and privacy and ensure secure data-sharing. Challenges also exist in building sustainable cross-sectoral partnerships to make the responsible use of such innovations possible for migration analysis and policy. These include access to data held by private entities and the setting up of collaborations that can facilitate the process; data reliability, as data from new sources are not representative of the population at large; incompatibility of concepts and definitions used for migration; limited analytical capabilities and unequal capacities across countries, which could exacerbate the digital divide; and finally, ethical and fundamental rights issues, particularly around the use of artificial intelligence-based systems for policymaking, protection of personal data, misuse of data for political purposes, and restriction of civil liberties.

Objectives and scope

Harnessing Data Innovation for Migration Policy: A Handbook for Practitioners responds to the need to strengthen capacities around migration data innovation, as voiced by policymakers and practitioners working in migration-related topics across several fields. It provides concise and accessible guidance to stakeholders interested in expanding the use of innovative data sources, tools and methods to complement traditional migration statistics. In particular, it aims to equip policymakers with the necessary tools for tapping into the abundance of existing data in the digital age, bridging practical and technical aspects of using data innovations, and explaining, step by step, how new data sources and methods can add value in relation to key policy-relevant migration questions.

As an example of such value, new data sources provide near-real-time observations. The timeliness and coverage of these sources can be particularly important in situations where traditional data-collection systems are disrupted, such as in contexts of global health or other kinds of crisis. During

[3] More information is available at https://data4migration.org/ and www.migrationdataportal.org/themes/big-data-migration-and-human-mobility (accessed 2 May 2022).

[4] More information is available at www.migrationdataportal.org/de/data-innovation.

the COVID-19 pandemic, for instance, Global Positioning System (GPS) data from Google Maps allowed governments to monitor the daily response to stay-at-home restrictions and their subsequent impact on infection rates. In disaster situations, the timely availability of data can critically impact peoples' lives. For instance, after the 2015 Nepal earthquake, mobile phone data helped monitor the location of affected populations and inform prompt humanitarian assistance.

In addition, new data sources offer fine-grained spatial resolution. Whether policymakers need information about public sentiment towards migrants, utilization rates of public spaces and roads, or population distributions in emergency camps, new data sources can support disaggregation by specific geographic locations, facilitating in-depth analyses of migration-related phenomena on all spatial levels, including cities, provinces and States. Traditional data sources can also potentially provide this level of detail; however, using surveys or adapting national censuses to measure migration on the local level would be much more costly and resource intensive, while most non-traditional data are already collected.

Some kinds of new data sources inherently reflect international patterns and are therefore well suited for studying international migration and human mobility. Companies such as Facebook, Google and Twitter naturally collect all data independently of national territorial borders. In fact, one of the first data innovation applications in migration used geotagged Internet Protocol (IP) addresses to estimate the number of people on the move worldwide. Furthermore, international news data collected by Google helped make short-term predictions of increasing immigration rates in Europe. While new data sources will not replace national censuses, household surveys and administrative records anytime soon, they can complement them and open new gateways for researching specific questions in particular contexts.

Data and new technologies can support programmes and policies related to migration across a range of priority policy issues – monitoring mobility induced by climate change and environmental degradation, and facilitating migration as an adaptation strategy; anticipating the spread of the current pandemic and preparing for future ones, ensuring "health-proof" mobility systems around the world; anticipating and effectively responding to humanitarian crises; and assisting migrants in vulnerable situations – along with broader migration policy topics, such as facilitating safe labour migration, monitoring public discourse and countering xenophobia and discrimination, ensuring access to information, and facilitating integration.

However, new ideas often come with old challenges. New data sources are particularly affected by three kinds of challenges. First, **reuse of private data and new technologies in the public policy sphere needs to earn public trust**. Unlike traditional data-collection methods, operated under clear informed-consent procedures and legal mandates, new kinds of data are generated automatically every time a person makes a phone call, interacts on social media or uses a GPS signal to navigate through traffic. This aspect induced many societies to become sceptical and wary towards the idea of sharing information via new data sources, especially against the backdrop of various high-profile incidents where data protection, security and privacy were violated.

A related challenge – which also represents an opportunity – is facilitating open discussions across sectors around trusted ethical and regulatory frameworks for reusing private data for research and policy purposes. New data sources are scattered across various entities, such as mobile network operators (MNOs), Internet companies, governments and national statistical offices, international organizations, and research institutions; and – even where political will to collaborate exists – such arrangements often come to a halt due to missing frameworks defining the dos and don'ts of data use. Developing such frameworks transparently and in line with ethical, privacy and security standards can contribute to earning public trust, enhancing the benefits of data reuse.

Finally, substantial technical capacities are needed, particularly in the public sector, to use data innovation at scale. New ideas and concepts can take shape only if the required technical skills and knowledge are available. Such capacities may often be in short supply in the public sector, unlike many businesses that have been competing for talent since the advent of the Internet.

The objective of this handbook is twofold: One is to raise awareness of the potential of using innovative data sources, tools and methods to complement traditional statistics, by providing a state-of-the-art account of existing innovations as well as governance aspects surrounding them. Another is to provide an accessible tool that can be used as a first step in activities aimed at enhancing data innovation capacities around the world.

The handbook provides practical information for a broad, non-technical audience of migration practitioners and policymakers working at the local, national and international levels – and anyone interested in data innovation applications to study human mobility and migration. It draws on the expertise of leading researchers and policymakers worldwide, featuring thematic applications of new data sources and methods, and reflecting on crucial governance, along with the ethical and security challenges associated with them. The handbook seeks to equip users with an accessible reference document that explains (a) how new data sources have added value in analysing different migration policy questions, relevant to the Sustainable Development Goals and the objectives of the Global Compact for Migration; (b) the main limitations and concerns to consider when using these data and methods; and (c) aspects of data governance needed to make innovations sustainable and beneficial for societies at large.

Defining data innovation

For the purpose of this handbook, the term "data innovation" will be used to refer to new data sources and innovative analytical methodologies offered by new technologies for the analysis of distinct policy-relevant migration issues. These can broadly be grouped under three key types of sources:

(a) Mobile phone data are de-identified call detail records of MNOs, which contain information about the time and associated cell tower of users' text messages, calls, and – in the case of extended data records – data exchanges. These data provide insights into human movements, particularly at the national and regional levels, in (almost) real time.

(b) Social media data are collected by social media services, such as Facebook, Twitter, Instagram, LinkedIn, TikTok and Google. They contain information on user activities – such as likes, posts, comments, shares, Internet searches, and use of hashtags and emojis – and user positions based on GPS location data, geotagged activities, Internet Protocol addresses and Wi-Fi network locations.

(c) Satellite data are images of the Earth taken from satellites around the world. They contain information about structures on the surface, such as roads, electricity lines, houses, destroyed infrastructures, and emergency camps. Satellite data can also refer to location data of GPS, geographic information systems and automated identification systems – for instance in the case of location data for ships and rescue vessels.

Data innovation also refers to innovative ways of analysing and combining data from traditional and non-traditional data sources, such as tools based on **artificial intelligence and machine learning**. Through mathematical models, artificial intelligence and machine learning-based systems can conduct automated analyses and collection of data relevant for migration policy and research without human intervention. Recent projects deployed artificial intelligence and machine learning to predict migration flows from specific countries, assess refugees' perception in destination communities, and support the analysis of conditions in refugee camps.

This handbook will demonstrate that non-traditional data sources, when used ethically and responsibly, can inform migration policy and programmes in their own right – and also when combined with traditional data sources.

Content and structure

The handbook is structured into two main sections. The first section showcases applications of new data sources and innovative methods offering insights relevant to different migration policy areas, such as estimating migrant stocks and flows, monitoring public sentiment towards migrants, and engaging migrant communities.

The first chapter, by Andrew J. Tatem, Claire A. Dooley, Shengjie Lai, Dorothea Woods, Alex Cunningham and Alessandro Sorichetta, highlights the importance of small-area mapping of population dynamics for policymaking, providing illustrative examples of integration of a variety of novel and traditional data sets.

The second chapter, by the United Nations Global Working Group on Big Data for Official Statistics, discusses the usefulness of various kinds of mobile positioning data (MPD) for migration statistics, outlining their various methodological approaches and limitations for measuring internal and international short- and long-term mobility.

The third chapter, by Jisu Kim, Emilio Zagheni and Ingmar Weber, discusses the applications and challenges of using social media data, particularly from Twitter and Facebook, to complement traditional statistics on migration.

Chapter 4 by Constantinos Melachrinos, Marcello Carammia and Teddy Wilkin provides an innovative framework for analysing current and anticipating future asylum-related migration.

Chapter 5 by Catherine Schneider, Rebeca Moreno Jimenez and Sofia Kyriazi provides a case study of the use of artificial intelligence-based predictive analytics to understand drivers of displacement, describing the process of setting up such a system, along with the limitations and lessons learned.

Chapter 6 by Jason Schachter, Angelica Menchaca and Antonio Bruce offers an innovative example of how the combination of traditional survey-based estimates and new data sources (airline passenger data) supported the estimation of net international migration in the United States of America.

Chapter 7 by Tuba Bircan, Albert Ali Salah and Alina Sîrbu focuses on the use of new data sources, including social media data and mobile phone data, to better understand migrant integration.

In Chapter 8, Francisco Rowe, Michael Mahony, Eduardo Graells-Garrido, Marzia Rango and Niklas Sievers present a case study of the use of Twitter data to monitor public sentiment towards migration across various countries.

Chapter 9 by Steffen Pötzschke focuses on how social networking sites – in this case Facebook and Instagram – can be used as cost-effective ways to sample migrants for research surveys.

Chapter 10 by Michael Newson and Niklas Sievers provides examples of the use of Google Analytics and onomastic analysis to map and study diaspora engagement.

Finally, Chapter 11 by Niklas Sievers and Marzia Rango summarizes the main kinds of challenges emerging from the thematic applications of new data sources and methods contained in the first section of the handbook.

The second section discusses the critical governance aspects related to new data sources and innovative analytical methods, including good practices for creating "data collaboratives" and appropriate regulatory frameworks to address ethical, security and data protection challenges.

In Chapter 12, David Reichel and Tamás Molnár provide guidance on the fundamental-rights-compliant use of new data sources and artificial intelligence in migration management.

Chapter 13, by Robert Trigwell, David Eduardo Zambrano and Gretchen Bueermann, lays out the ethical foundations and principles for the use of advanced data science methods in contexts of humanitarian emergencies, as developed by the Humanitarian Data Science and Ethics Group.

Chapter 14, by Miguel Luengo-Oroz, Katherine Hoffmann Pham and Rebeca Moreno Jiménez, discusses the development of predictive modelling and awareness tools in situations of uncertainty, describing a three-part decision support system designed to help anticipate human displacement, and stressing the importance of interdisciplinary approaches and partnerships.

In Chapter 15, Michele Vespe, Stefano Maria Iacus, Umberto Minora, Carlos Santamaria, Francesco Sermi, Spyridon Spyratos and Dario Tarchi offer insights from a business-to-government initiative between the European Commission and various MNOs in Europe to help monitor the impact of COVID-19-related restrictions on intra-European mobility, highlighting aspects in the areas of security, privacy, fundamental rights and commercial confidentiality.

Finally, Chapter 16 by Stefaan G. Verhulst and Andrew Young discusses the need for and kinds of data collaboratives – an emergent form of public–private partnerships centred around data reuse across sectors and actors – providing step-by-step guidance to establish such partnerships for migration.

I

HARNESSING DATA INNOVATION FOR MIGRATION POLICY: THEMATIC APPLICATIONS

MEASURING AND FORECASTING MIGRATION STOCKS AND FLOWS

GEOSPATIAL DATA INTEGRATION TO CAPTURE SMALL-AREA POPULATION DYNAMICS

Andrew J. Tatem,[1] *Claire A. Dooley,*[2] *Shengjie Lai,*[3] *Dorothea Woods,*[4] *Alex Cunningham*[5] *and Alessandro Sorichetta*[6]

Summary

In this chapter, we highlight the importance of small-area data on population distributions for supporting policymaking. We emphasize how population distributions vary in different ways at different spatial and temporal scales. Various "big" data sets now exist to capture some of these dynamics, each with their own strengths, but also many drawbacks. We discuss how harmonizing and integrating data sets into a common geospatial framework enables the strengths of different data sets representing features of mobility and migration to be brought together, building on each other. We provide an overview of data sets and methods for such integration, then present three illustrative case studies where such integration has been used to support decision-making.

Background

Data on population counts for small areas underlie almost all areas of governance, policymaking and resource allocation. Knowing accurately how many people reside in an area in a certain time period is important for efficient response to natural disasters, ensuring sufficient aid is delivered equitably and that representation in parliaments is fair. Without such data, children can be missed in vaccination programs, election planning becomes challenging, and infrastructure development does not meet needs. However,

[1] Andrew J. Tatem is Spatial Demography and Epidemiology Professor at the University of Southampton. He is Founder and Director of WorldPop (www.worldpop.org), an applied research group focused on the integration of geospatial data for mapping population distributions, demographics and dynamics at small-area scales.

[2] Claire A. Dooley is a quantitative researcher working in the area of demography and health. Her work involves the development and application of spatial methods to understand displacement patterns and estimate health indicators for populations impacted by conflict and protracted crises.

[3] Shengjie Lai has been long engaged in interdisciplinary research on human mobility, environmental changes and epidemics, to provide an improved evidence base for development and infectious disease control decision-making. His work has been published in scientific journals such as *Nature*, *Science*, the *BMJ*, the *Lancet Infectious Diseases* and *Nature Human Behaviour*.

[4] Dorothea Woods is a geographic information systems trainer and analyst who uses a range of methods to add a spatial component and gain new insights into projects in the ecology and social sciences field. Her key interest lies in migration and human mobility, where she works with novel data sets and big data.

[5] Alex Cunningham has a background in geographic information systems and remote sensing, and has spent recent years focused on identifying, processing and generating key statistics from a wide range of spatio-temporal data sets to aid projects exploring seasonal population movements.

[6] Alessandro Sorichetta is Associate Professor in Applied Geospatial Research at WorldPop, University of Southampton. His research in the field of spatial demography focuses on mapping the global distribution of the human population, including its demographic and socioeconomic characteristics, and modelling human migration and mobility.

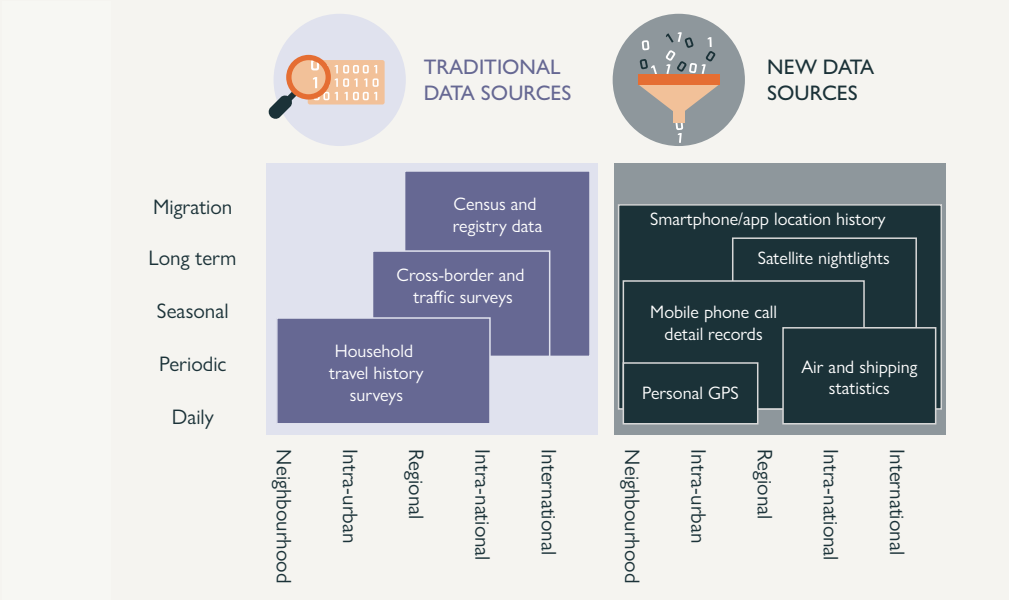
population numbers change substantially over time and space, and regular enumerations at small-area scales can be prohibitively expensive and logistically challenging.

National population and housing censuses typically provide the most comprehensive, complete and accurate source of data on residential population numbers, characteristics and some information on residence changes over the prior 1-to-5-year period. Such data-collection exercises are typically the largest peacetime operations that governments engage in and therefore are normally undertaken just once every decade. In some resource-poor settings, the gaps between such data-collection efforts can be many decades. Between these enumerations, population distributions can change substantially – for example, through daily commutes, seasonal migrations or longer-term displacement. The irregularity of census-based enumeration means that these changes are generally not captured, and estimating them accurately for small areas through standard projection methods is challenging. To capture such population dynamics at small-area levels therefore requires consideration of alternative data sources that can more regularly register changes.

Data sets to support small-area mapping of population dynamics

The past decade has seen the rise of novel, big data sets that capture aspects of population dynamics at a level of richness previously unavailable. Figure 1 highlights the spatial scales and temporal frequencies of population movements that "traditional" and "novel" methods of data collection cover. Each has its own set of weaknesses and biases, and it is often only through linking data sets that some of these weaknesses are mitigated, as they build on the strengths of each other. An overview of these different data sources is provided here; a more detailed description of each can be found in the work of Tatem (2014).

Figure 1. Sources of data for measuring population mobility



Note: Sources of data for measuring population mobility covering temporal (y-axis) and spatial (x-axis) scales of movement, broken down by (a) traditional sources and (b) novel or big data sources. The plot highlights how novel forms of data cover a wider range of spatial scales and temporal frequencies, particularly those derived from mobile phones. (Adapted from: Tatem, 2014)

Traditional sources of data

Figure 1 highlights that methods and data sets for capturing human mobility information from neighbourhood to international scales and from daily to long-term frequencies have been available for decades. However, these methods and the resulting data are typically undertaken only irregularly and can be expensive. National population and housing censuses have been implemented by most countries every decade for half a century, and they often represent the largest and most expensive peacetime operations that governments have undertaken. They typically provide data on changes in residence over the 1-to-5-year period prior to the census. Such data are incredibly valuable for understanding domestic and international migration patterns at small spatial scales and demographics. Their implementation just once a decade (sometimes much less frequent in resource-poor settings) means that data for the intervening years are lacking and little data on shorter-term movements are collected. Registries and administrative data can fill these gaps, but data-collection systems are incomplete or lacking entirely for many countries, particularly across low-income regions such as sub-Saharan Africa.

Surveys continue to play a vital role in our measurement and understanding of population mobility. These can be geographically targeted at specific points of interest, such as border-crossing points, airports or other high-traffic points to obtain data on flows and the characteristics of travellers. Moreover, household surveys often provide a valuable source of population-level information on mobility patterns, together with demographic breakdowns and a range of other factors, including motivation for travel. The major limitation of survey data, however, is the expense in collection, meaning that rich data exist for specific time points and locations, but regular collection over vast geographic areas and populations is difficult. Moreover, designing sampling methods to ensure that migrants, nomads and other mobile populations are well represented can be challenging.

Novel sources of data

A much wider range of spatial scales and temporal frequencies of movements are captured by the novel (or often termed "big") data sets shown in Figure 1 – such as those from mobile phones, satellite imagery and air-traffic statistics. Some suffer from the same limitations as surveys, particularly the cost and logistics of producing large-area insights over long time periods, such as the use of wearable GPS trackers (Vazquez-Prokopec et al., 2009). However, some benefits are clear, including overcoming some survey limitations, such as recall bias, to provide detailed insights into individual mobility patterns (Floyd et al., 2020; Searle et al., 2017). Air and shipping statistics share some similarities with cross-border and traffic surveys. Still, today big data sets on individual air-travel tickets or GPS tracks of each ship globally are available, forming incredibly rich information on population movements across the world. These register the movements of individuals taking certain forms of transport, but the measurement stops once an individual leaves a port or airport. One form of data that has shown value recently in providing a surrogate for seasonal variations in urban populations in resource-poor settings is night-time satellite images. Across the world, movements can be highly seasonal due to holidays or seasonal labour migration. In the Sahel region of West Africa, this is especially the case in the dry season, when thousands of agricultural labourers move to cities to find alternative work. The switching on of lights and lighting of fires are visible in night-time images and have proven to be a valuable indicator of timings and magnitudes of population mobility and migration in the absence of other data (Bharti et al., 2011, 2016).

The largest area of Figure 1 covered by individual data sources derives from the growing use of mobile devices globally. Call detail records (CDRs) are maintained by mobile phone network operators for billing purposes. They include a phone device ID, the time of communications sent or received, and the cell tower that the communications were routed through. With regular communications to/from a phone and cell tower locations, the approximate movement of the phone can be mapped. Across the millions of phone users, such data have provided a richness of mobility metrics never seen before, covering entire countries at fine spatial and temporal scales. Significant biases exist, with children, the poor, the elderly and some geographical areas often poorly represented, and cross-border measurements typically not possible; but valuable insights can often be obtained despite these (Wesolowski et al., 2013b). CDRs, however, often remain difficult to access, with legally binding agreements with network operators typically required; also, complex data access set-ups are at times necessary due to the sensitivity of such data and the commercial value to the operators. Finally, the world's shift towards smartphones means that GPS location histories obtained by tech companies and app developers are becoming increasingly representative of movement patterns as they capture growing proportions of the population. The value of Google Location History data for capturing mobility across multiple countries over long periods of time has been demonstrated (Ruktanonchai et al., 2018; Kraemer et al., 2020), while similar aggregated and anonymized data sets are being made available by Facebook and Apple to support COVID-19 response efforts.

Geospatial data integration

Each of the data sets outlined above and shown in Figure 1 is typically collected independently, with the comparison and integration with other forms of mobility data often not explicitly planned. Consequently, these data come in all shapes and sizes, making their integration challenging. However, one common feature of all of them is geographical information. The movements captured can generally be mapped to specific administrative units, towns, villages or even precise GPS locations. This provides a framework to overlay and link data sets together, enabling comparisons as well as the development of analyses and models that draw on insights from multiple types of mobility data. Significant effort and care are often required to ensure comparability, but multiple studies have endeavoured to uncover relationships and strengthen mobility insights through linking data sets. These include census migration and CDRs (Wesolowski et al., 2013a), smartphone location history and travel history surveys (Ruktanonchai et al., 2018), and CDRs and travel history surveys (Wesolowski et al., 2014). The following sections dive more deeply into three examples of the integration of geospatially referenced mobility data with other forms of geospatial data to tackle specific policy-relevant challenges: (a) the integration of population displacement surveys with census and satellite data to map contemporary population distributions; (b) linking mobile phone CDRs with disease transmission suitability mapping to prioritize intervention delivery; (c) developing spatial disease spread models driven by smartphone location histories and CDRs to guide policy in outbreak response.

Case studies

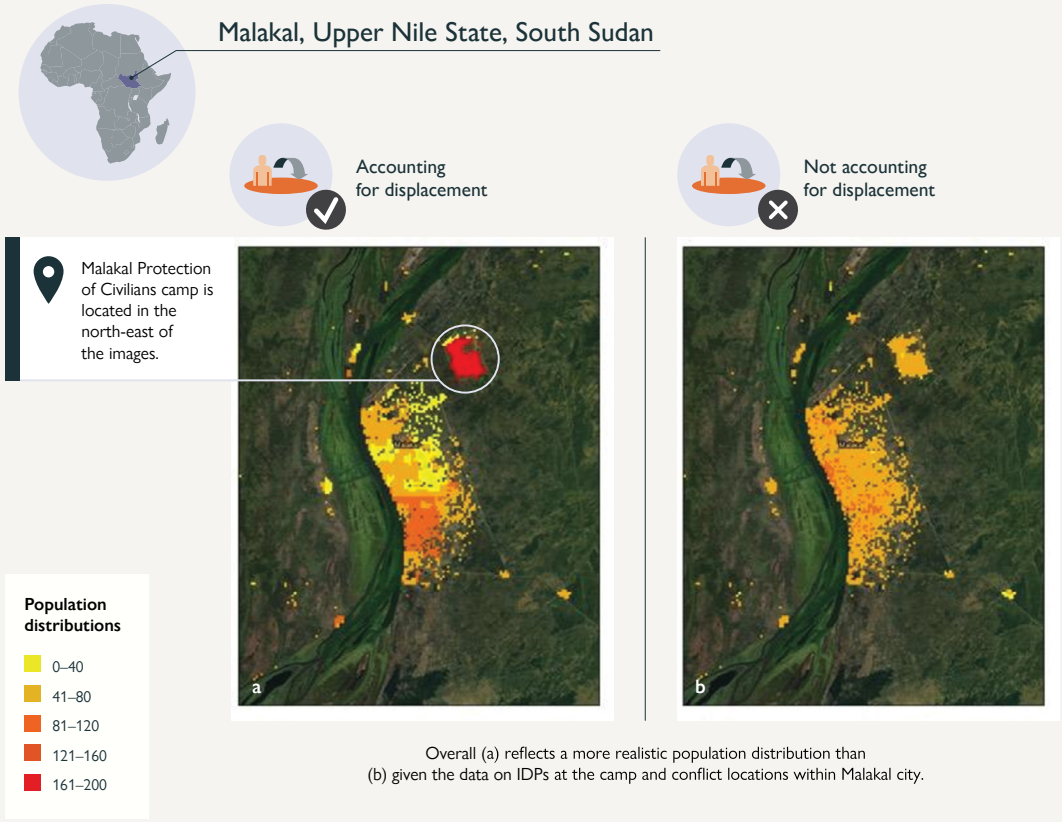
Accounting for displacement in population mapping in South Sudan

South Sudan's last census was conducted in 2008, prior to its independence from the Sudan in 2011. Based on these population counts, South Sudan's National Bureau of Statistics (2015) produces subnational population projections using fertility and mortality rates. However, the population projections do not account for population movement due to the lack of validated migration data. Not accounting for movement is incredibly problematic because since independence, South Sudan has experienced widespread conflict that has led to the displacement of many people across the country as well as into neighbouring territories. Additionally, annual flooding events are common and, depending on their magnitude, often cause displacement. Round 9 of IOM's Displacement Tracking Matrix (DTM) (carried out between July and September 2020) reports that there are approximately 1.6 million internally displaced persons (IDPs) across South Sudan, and the Office of the United Nations High Commissioner for Refugees (UNHCR) reports that there are approximately 2.2 million South Sudanese refugees residing in Uganda (40.4%), the Sudan (33.3%), Ethiopia (16.6%), Kenya (5.6%) and the Democratic Republic of the Congo (4.1%), as of October 2020 (IOM, 2020, 2021; UNHCR, 2020).

To generate an accurate population distribution that accounts for displacement, it is necessary to integrate a number of different displacement-related data sources. In recent years, the spatial coverage of IOM's DTM has expanded, with the most recent rounds including a large proportion of the country. The DTM data contain geolocated destinations of IDPs as well as the county (administrative level 2) from which the majority of IDPs at each destination location have been displaced. The UNHCR's data portal does not provide any information about the subnational place of origin of refugees; however, the Regional Intention Survey includes the approximate proportion of refugees from each state (administrative level 1) within South Sudan based on sampled households across 15 refugee camps: the Democratic Republic of the Congo (3 camps), Ethiopia (3 camps), Kenya (3 camps), Uganda (3 camps), the Sudan (2 camps) and the Central African Republic (1 camp) (UNHCR, 2019). IOM's data on destinations of IDPs can be used directly to map displaced populations and adjust census projections to reflect the corresponding areas of increased population size. The challenge comes when attempting to map the locations of depleted population sizes compared to the census projections. From IOM and UNHCR data on place of origin by administrative units, disaggregation approaches can be applied to infer estimated numbers of people displaced from all areas of the country at a high spatial resolution. For disaggregation to be effective and accurate, appropriate predictors of place of origin are needed. For South Sudan, the key drivers of displacement are conflict and flooding, and therefore variables such as distance to fatal conflict events and precipitation levels during rainy seasons are required. These types of variables have been derived by applying geocomputing methods to spatial data sets such as the Armed Conflict Location and Event Data Project (ACLED) (Raleigh et al., 2010) and WorldClim's monthly weather measurements (Fick and Hijmans, 2017). The resulting disaggregation using these geospatial variables successfully infers larger-scale displacement from locations of intense conflict and flooding.

With estimates of displacement mapped by origin and destination, subnational census projections can be adjusted to produce a more realistic population distribution. The 2019 South Sudan population data set produced by WorldPop (2020) (as part of the GRID3 project)[7] using this approach, integrated with building footprints mapped from recent satellite imagery, provides grid cell-level (~100m x 100m) population estimates that account for displacement (Ecopia AI and Maxar Technologies, 2020; Dooley et al., 2020, 2021a, 2021b). The high spatial resolution of this data set allows the detailed level of information about numbers and locations of IDPs to be explicitly included and not lost through aggregation to larger spatial units, while still providing flexibility to data users to perform aggregations for their own spatial units – e.g. health catchments, settlements. In Figure 2, we highlight the added benefit of including displacement in population estimates by comparing the GRID3 2020 South Sudan estimates to disaggregated census data that do not account for displacement.

Figure 2. Population distribution differences accounting for displacement



Overall (a) reflects a more realistic population distribution than (b) given the data on IDPs at the camp and conflict locations within Malakal city.

Note: A comparison of two population data sets in Malakal, Upper Nile State, South Sudan. The Malakal Protection of Civilians camp is located in the northeast of the images. The maps show disaggregated census projections that (a) account for displacement (Dooley et al., 2021a) and (b) do not account for displacement (Bondarenko et al., 2020). Overall, Figure 2a reflects a more realistic population distribution than Figure 2b, given the data on IDPs at the camp and conflict locations within the city of Malakal. (Service layer credits: Esri, DigitalGlobe, GeoEye-1, Earthstar Geographics, French Space Agency (CNES)/Airbus DS, United States Department of Agriculture, United States Geological Survey, AeroGRID, Institut national de l'information géographique et forestière (IGN)).

These maps are for illustration purposes only. The boundaries and names shown and the designations used on these maps do not imply official endorsement or acceptance by the International Organization for Migration.

[7] More information is available at www.grid3.org.

Mapping population mobility in Namibia to support malaria-elimination efforts

The world has made great strides towards malaria eradication over recent decades. Many countries on the edge of the disease's endemic range have achieved national elimination through a combination of interventions, together with the impact of factors such as urbanization and poverty alleviation. Namibia in Southern Africa is one country that has made significant progress towards malaria elimination over the past two decades, with annual case numbers dropping from a level of hundreds of thousands to around or below the ten-thousand mark.

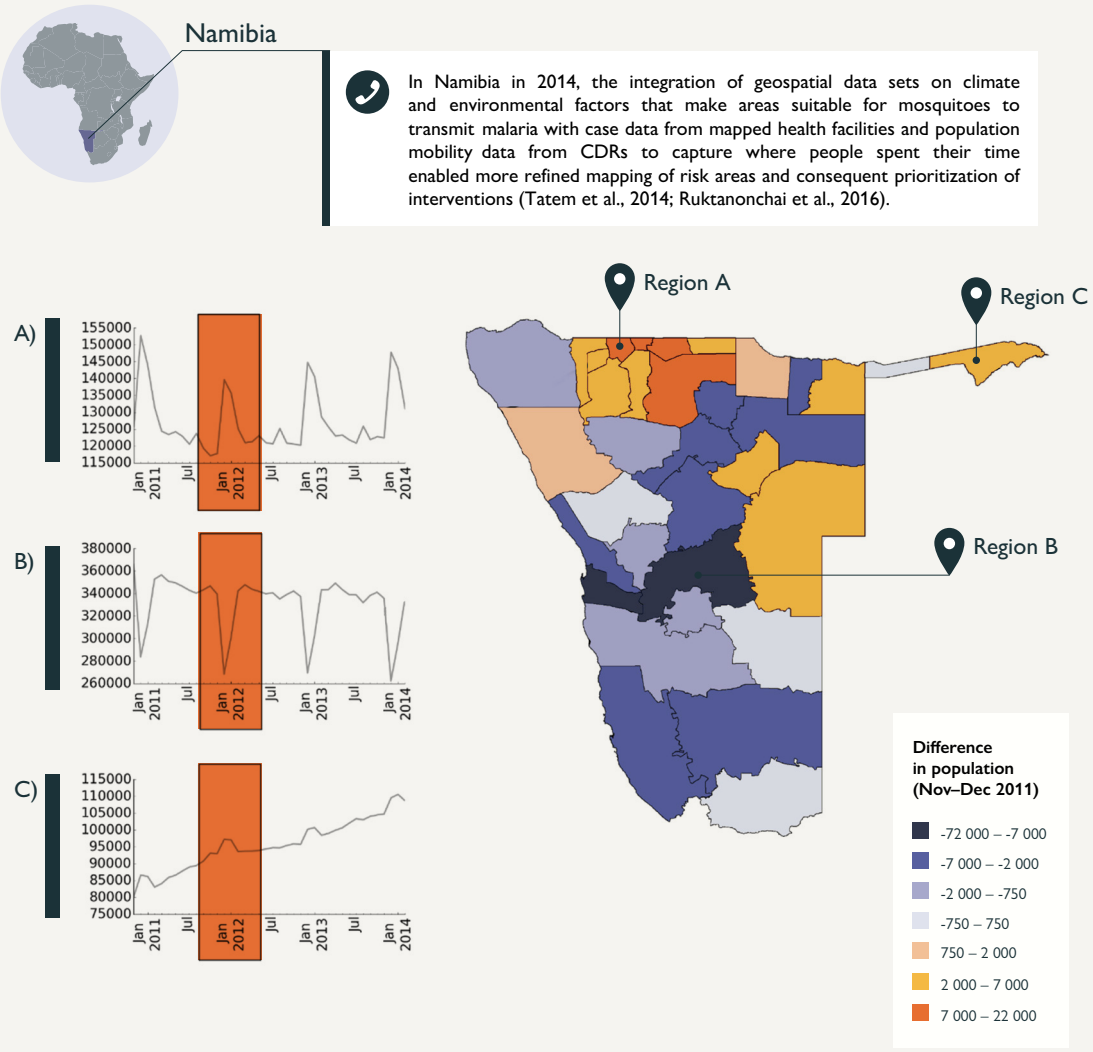
Once malaria case numbers become small and a country is aiming for elimination of the disease, the importation of infections and targeting of interventions become high priorities. Malaria transmission typically occurs in places where densities of mosquitoes are high enough, and where these interact with a population in evenings/night-time, when biting occurs. Surveillance data from health facilities may highlight clusters of cases, but those infections may have resulted from being bitten by infectious mosquitoes far from the facility. Understanding where the transmission took place can help identify hotspots of transmission and, therefore, tailor interventions accordingly to reach elimination. Mobility data can therefore play an important role in understanding one component of this – where people spend their time.

In Namibia in 2014, the integration of geospatial data sets on climate and environmental factors that make areas suitable for mosquitoes to transmit malaria, with case data from mapped health facilities and population mobility data from CDRs, to capture where people spent their time, enabled more refined mapping of risk areas and consequent prioritization of interventions (Tatem et al., 2014; Ruktanonchai et al., 2016). The maps showed that some areas of the country only saw cases because of a high proportion of importations from higher transmission areas. Targeting these higher transmission areas would likely have a disproportionate effect on overall case numbers and be a more effective use of limited resources. The maps were used by the National Vector Borne Disease Control Programme to prioritize bed net distribution to the areas and people most likely important to the transmission cycle.[8]

Following requests from the World Health Organization (WHO) office in Namibia, the value of the CDRs was further demonstrated through integration with census data in providing estimates of monthly changes in population numbers across the country (zu Erbach-Schoenberg et al., 2016; see also Figure 3). When linked with maps of health facility catchments, estimates of seasonal changes and demands for health interventions were obtained, as well as new denominators for health metrics. Finally, the strong correlations between census-based migration flows and phone-based estimates demonstrated the value of CDRs for more regular updates to national migration statistics (Lai et al., 2019).

[8] More information is available at https://dataimpacts.org/project/malaria/.

Figure 3. Mapping seasonal population changes using mobile phone data



Source: Adapted from: zu Erbach-Schoenberg et al., 2016.

Note: Estimates of seasonal changes in population numbers by region in Namibia based on the integration of census and mobile phone CDRs. The map shows estimated differences in population for health districts between November and December 2011. The inset graphs show predicted population numbers for selected health districts over the January 2011–January 2014 period.

This map is for illustration purposes only. The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by the International Organization for Migration.

Incorporating mobility data into COVID-19 transmission modelling

Modern transportation plays a key role in the long-distance and rapid spread of infectious diseases due to the high mobility of human hosts across the globe. In the early stages of the COVID-19 pandemic, there was an urgent need to understand the spread risk and geographic range of COVID-19 transmission, as well as the effectiveness of non-pharmaceutical interventions (NPIs) on COVID-19 containment and mitigation. As it was an emerging disease, vaccines and drugs were not expected to be available in a short period. NPIs – such as case identification and isolation, contact tracing,

travel restrictions, physical distancing, face masking, hand-washing, school and workplace closures, and even lockdown (cordon sanitaire) of cities or countries – were implemented across the world. These aimed to reduce transmission, infections and deaths, thereby delaying the epidemic peak, and buying time for health-care preparations and vaccines to be used later on. Prior to COVID-19, however, few studies had systematically investigated the effectiveness of NPIs on various infectious diseases as well as how they should be implemented across space and time to mitigate a pandemic's negative effects and protect vulnerable populations at risk of severe outcomes. The lack of relevant evidence led to delays in COVID-19 containment in the early stages of the outbreak, and uncoordinated interventions also reduced the effectiveness of these measures to prevent resurgences. Aggregated and anonymized location history data from smartphones and mobile phone CDRs, however, have provided a vital basis for studies assessing the effectiveness of NPIs in containing COVID-19 transmission (Lai et al., 2020; Ruktanonchai et al., 2020).

A travel network-based epidemiological model was built to estimate the numbers of susceptible, exposed, infectious and recovered/removed subpopulations per day within defined geographical areas in China, as well as the number of infected travellers moving between each pair of study areas. Based on historical and near real-time anonymized and geographically aggregated human mobility data obtained from smartphone users of Baidu's location-based services, and data on delay from illness onset to reporting of cases across the country, the modelling framework was used to reconstruct the transmission dynamics of COVID-19 across 340 prefecture-level cities in mainland China from 1 December 2019 up until 30 April 2020. Moreover, comparable before-and-after analyses were conducted to quantify the relative effect of the three major groups of NPIs: (a) the restriction of intercity population movement, (b) the identification and isolation of cases, and (c) the reduction of inner-city travel and contact to increase social distance (Lai et al., 2020). Additionally, mobility metrics across Europe during the pandemic were derived from anonymized and aggregated Google Location History data, along with Vodafone CDRs, to illustrate how coordinated exit strategies could delay continental resurgence and limit COVID-19 community transmission (Ruktanonchai et al., 2020).

The mobility data were vital to understanding the dynamics of spread, especially in China, where the outbreak coincided with the Lunar New Year holiday (Figure 4), which drives the largest annual movement of people on the planet. The studies estimated that the COVID-19 outbreak in China increased exponentially prior to the Lunar New Year, but the peak of epidemics across the country quickly appeared during the Lunar New Year holiday. Without NPIs, the COVID-19 cases in mainland China would likely have shown a 51-fold (interquartile range 33–71) increase in Wuhan, a 92-fold (58–133) increase in other cities in Hubei, and a 125-fold (77–180) increase in other provinces by 29 February 2020 (Lai et al., 2020). However, the lockdown of Wuhan might not have prevented the seeding of the virus from the city, as the travel ban was put in place at the latter stages of the pre-Lunar New Year population movement out of the city. Nevertheless, if intercity travel restrictions were not implemented, cities and provinces outside of Wuhan would have received more cases from the city, and the affected geographic range would have expanded to the remote western areas of China.

Figure 4. Estimated travel flows in China with and without restrictions at the start of the COVID-19 pandemic

The effectiveness of different interventions varied. Generally, the early detection and isolation of cases were estimated to quickly and substantially prevent more infections than contact reduction and social distancing across the country (5-fold versus 2.6-fold). However, without the intervention of contact reductions, the epidemics would increase exponentially across regions in the longer term. Therefore, combined NPIs achieved the strongest and most rapid effects in terms of COVID-19 outbreak containment. Additionally, these studies also suggested that a resurgent continental epidemic could occur as many as five weeks earlier when well-connected countries with existing stringent interventions end those interventions prematurely. Appropriate coordination of NPIs and vaccine rollouts could significantly improve the likelihood of containing community transmission throughout Europe and prevent resurgences. In particular, synchronizing intermittent lockdowns across Europe meant that half as many lockdown periods were required to end community transmission continent-wide (Ruktanonchai et al., 2020). These studies show that models using mobility data derived from novel, big data sources have significantly improved our understanding of this pandemic and intervention efficacy. A combined, coordinated and timely NPI strategy could substantially reduce

COVID-19 transmission across countries to avoid resurgence, and each study formed part of the evidence base used to guide actions taken by organizations such as the Chinese Center for Disease Control and Prevention (China CDC) and the European Centre for Disease Prevention and Control (ECDC). Given the improving access to timely anonymized population movement data for supporting COVID-19 mitigation across the globe, the potential exists to monitor and assess the effectiveness of NPIs to inform strategies against future COVID-19 waves and potential future pandemics.

Discussion

The value of accurately quantifying population distributions and dynamics at small-area scales is clear. Just focusing on one vital area – as an example, health monitoring and interventions around child mortality in low-income settings – in terms of denominators for health metrics or birth and death registration, it is hard to know whether a national deworming programme for children in Sierra Leone or a vaccination programme for pertussis in Nigeria is reducing mortality when less than 4 per cent of deaths are registered. However, even if 100 per cent of deaths are registered, it remains challenging to implement the programmes and place the number of deaths in context without reliable multi-temporal, disaggregated data on population numbers and distributions, particularly when seasonal dynamics are strong and population groups are highly mobile. Careful surveillance can guide public health in a country and track disease outbreaks that could spread beyond borders. Improving detection and measurement of the numerator without attention to the denominator, however, risks providing an inaccurate picture. The analyses outlined in the Namibia case study showed that improved quantification of seasonal variations in denominator populations changed malaria incidence measures by over 30 per cent (zu Erbach-Schoenberg et al., 2016). Moreover, the reliance on static and ageing figures for denominators leads to the common occurrence of 200 per cent vaccination rates, and/or incidence measures fluctuating by season where population mobility is high (Cutts et al., 2016).

The case studies outlined above highlight how novel and big data sets that capture aspects of human mobility can fill vital data gaps to support decision-making. In each case, the mobility data provide one component of geospatial modelling efforts that draw on multiple sources, building on the strength of each to attempt to overcome gaps, biases and weaknesses. Each data set comes with its own sensitivities, and ethics review board assessments of proposed analyses are vital to ensure the protection of individuals – a single data set may comply with a predetermined standard for anonymization, yet when linked with other geospatially referenced data sets, this can open up risks of reidentification. Understanding and quantifying biases in new forms of data represent an ongoing area of research, and intercomparisons with more traditional forms of data remain vital to understand and account for uncertainties. The new forms of mobility data should not be seen as a replacement for traditional sources – they are more of a complement – and continued investment in census, registry and survey data collection remains vital to anchor insights from big data in reality.

REFERENCES*

Bharti, N., A. Djibo, A.J. Tatem, B.T. Grenfell and M.J. Ferrari
2016 Measuring populations to improve vaccination coverage. *Scientific Reports,* 6.

Bharti, N., A.J. Tatem, M.J. Ferrari, R.F. Grais, A. Djibo and B.T. Grenfell
2011 Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery. *Science*, 334(6061):1424–1427.

Bondarenko, M., D. Kerr, A. Sorichetta and A.J. Tatem
2020 Census/projection-disaggregated gridded population datasets, adjusted to match the corresponding UNPD 2020 estimates, for 51 countries across sub-Saharan Africa using building footprints. WorldPop, University of Southampton.

Cutts, F.T., P. Claquin, M.C. Danovaro-Holliday and D.A. Rhoda
2016 Monitoring vaccination coverage: Defining the role of surveys. *Vaccine*, 34(35):4103–4109.

Dooley, C.A., G. Boo, D.R. Leasure and A.J. Tatem
2020 Gridded maps of building patterns throughout sub Saharan Africa, version 1.1. WorldPop, University of Southampton.

Dooley, C.A., W.C. Jochem, D.R. Leasure, A. Sorichetta, A.N. Lazar and A.J. Tatem
2021a South Sudan 2020 gridded population estimates from census projections adjusted for displacement, version 2.0. WorldPop, University of Southampton.

Dooley, C.A., W.C. Jochem, A. Sorichetta, A.N. Lazar and A.J. Tatem
2021b Description of methods for South Sudan 2020 gridded population estimates from census projections adjusted for displacement, version 2.0. WorldPop, University of Southampton.

Ecopia AI and Maxar Technologies
2020 Digitize Africa data.

Fick, S.E. and R.J. Hijmans
2017 WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Floyd, J.R., J. Ogola, E.M. Fèvre, N. Wardrop, A.J. Tatem and N.W. Ruktanonchai
2020 Activity-specific mobility of adults in a rural region of western Kenya. *PeerJ*.

** All hyperlinks were working at the time of writing this report.*

International Organization for Migration (IOM)

2020 South Sudan – Baseline Locations Round 9. Displacement Tracking Matrix (DTM). Available at https://displacement.iom.int/datasets/south-sudan-baseline-locations-round-9.

2021 South Sudan – Baseline Assessment Round 9 – IDP and Returnee. DTM. Available at https://displacement.iom.int/datasets/south-sudan-baseline-assessment-round-9-idp-and-returnee.

Kraemer, M.U.G., A. Sadilek, Q. Zhang, N.A. Marchal, G. Tuli, E.L. Cohn, Y. Hswen, T.A. Perkins, D.L. Smith, R.C. Reiner Jr and J.S. Brownstein

2020 Mapping global variation in human mobility. *Nature Human Behaviour*, 4:800–810.

Lai, S., N.W. Ruktanonchai, L. Zhou, O. Prosper, W. Luo, J.R. Floyd, A. Wesolowski, M. Santillana, C. Zhang, X. Du, H. Yu and A.J. Tatem

2020 Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature*, 585:420–413.

Lai, S., E. zu Erbach-Schoenberg, C. Pezzulo, N.W. Ruktanonchai, A. Sorichetta, J. Steele, T. Li, C.A. Dooley and A.J. Tatem

2019 Exploring the use of mobile phone data for national migration statistics. *Palgrave Communications*, 5.

National Bureau of Statistics

2015 Population projections for South Sudan by county from 2015–2020. Available at www.ssnbss.org/home/document/census/population-projections-for-south-sudan-by-county-from-2015-to-2020/.

Office of the United Nations High Commissioner for Refugees (UNHCR)

2019 Regional Intention Survey of South Sudanese Refugees: Central African Republic, Democratic Republic of the Congo, Ethiopia, Kenya, Sudan, Uganda – June 2019. Available at https://microdata.unhcr.org/index.php/catalog/224/download/674.

2020 Regional overview of the South Sudanese refugee population: 2020 South Sudan Regional RRRP. Available at https://data2.unhcr.org/en/documents/details/79631.

Raleigh, C., A. Linke, H. Hegre and J. Karlsen

2010 Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature. *Journal of Peace Research*, 47(5):651–660.

Ruktanonchai, N.W., P. DeLeenheer, A.J. Tatem, V.A. Alegana, T.T. Caughlin, E. zu Erbach-Schoenberg, C. Lourenço, C.W. Ruktanonchai and D.L. Smith

2016 Identifying malaria transmission foci for elimination using human mobility data. *PLOS Computational Biology*, 12(4).

Ruktanonchai, N.W., J.R. Floyd, S. Lai, C.W. Ruktanonchai, A. Sadilek, P. Rente-Lourenco, X. Ben, A. Carioli, J. Gwinn, J.E. Steele, O. Prosper, A. Schneider, A. Oplinger, P. Eastham and A.J. Tatem
2020 Assessing the impact of coordinated COVID-19 exit strategies across Europe. *Science*, 369(6510):1465–1470.

Ruktanonchai, N.W., C.W. Ruktanonchai, J.R. Floyd and A.J. Tatem
2018 Using Google Location History data to quantify fine-scale human mobility. *International Journal of Health Geographics*, 17.

Searle, K.M., J. Lubinda, H. Hamapumbu, T.M. Shields, F.C. Curriero, D.L. Smith, P.E. Thuma and W.J. Moss
2017 Characterizing and quantifying human movement patterns using GPS data loggers in an area approaching malaria elimination in rural southern Zambia. *Royal Society Open Science*, 4(5).

Tatem, A.J.
2014 Mapping population and pathogen movements. *International Health*, 6(1):5–11.

Tatem, A.J., Z. Huang, C. Narib, U. Kumar, D. Kandula, D.K. Pindolia, D.L. Smith, J.M. Cohen, B. Graupe, P. Uusiku and C. Lourenço
2014 Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malaria Journal*, 13.

Vazquez-Prokopec, G.M., S.T. Stoddard, V. Paz-Soldan, A.C. Morrison, J.P. Elder, T.J. Kochel, T.W. Scott and U. Kitron
2009 Usefulness of commercially available GPS data-loggers for tracking human movement and exposure to dengue virus. *International Journal of Health Geographics*, 8.

Wesolowski, A., C.O. Buckee, D.K. Pindolia, N. Eagle, D.L. Smith, A.J. Garcia and A.J. Tatem
2013a The use of census migration data to approximate human movement patterns across temporal scales. *PLOS ONE*, 8(1).

Wesolowski, A., N. Eagle, A.M. Noor, R.W. Snow and C.O. Buckee
2013b The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface*, 10(81).

Wesolowski, A., G. Stresman, N. Eagle, J. Stevenson, C. Owaga, E. Marube, T. Bousema, C. Drakeley, J. Cox and C.O. Buckee
2014 Quantifying travel behavior for infectious disease research: A comparison of data from surveys and mobile phones. *Scientific Reports*, 4.

WorldPop

2020 South Sudan 2019 gridded population estimates from census projections adjusted for displacement, version 1.0. School of Geography and Environmental Science, University of Southampton. Available at https://eprints.soton.ac.uk/437804/.

zu Erbach-Schoenberg, E., V.A. Alegana, A. Sorichetta, C. Linard, C. Lourenço, N.W. Ruktanonchai, B. Graupe, T.J. Bird, C. Pezzulo, A. Wesolowski and A.J. Tatem

2016 Dynamic denominators: The impact of seasonally varying population numbers on disease incidence estimates. *Population Health Metrics*, 14.

ADDITIONAL READING

Wardrop, N.A., W.C. Jochem, T.J. Bird, H.R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman and A.J. Tatem

2018 Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences of the United States of America*, 115(14):3529–3537.

HARNESSING DATA FROM MOBILE NETWORK OPERATORS FOR MIGRATION STATISTICS

United Nations Global Working Group on Big Data for Official Statistics

Big data sources for migration

Among global political circles, migration is considered a high-priority policy issue. Despite this, only a fraction of the official statistics captures the complexity of migration as a phenomenon, challenging the effectiveness of policymaking on this issue. Timely, reliable and comparable data on migration are key for evidence-informed policymaking. At the same time, traditional data sources (such as surveys and censuses) struggle to capture the dynamics of modern migration, including forced migration due to conflicts and environmental threats, migration of highly skilled labour, and mobile and hidden populations.[1] Big data remain a novel yet largely untapped source of information that could provide valuable real-time insights into migration. What is more, traditional data-collection methods cannot fully measure the mobility of people. By tapping into big data sources, such as mobile positioning and social media data, it is possible to fill the gaps left in currently available statistics.

There are several potential sources of big data, which can broadly be grouped into three categories (Global Migration Group, 2017):

(a) Sensor based (e.g. Earth observation data from satellite imagery)
(b) Internet based (e.g. social media)
(c) Mobile phone based (e.g. call detail records (CDRs) or mobile positioning data (MPD))

Each of these sources holds its own value. However, their limitations should be recognized. Sensor-based data are particularly useful for assessing migration due to environmental threats and, as such, can provide good background knowledge for explaining migration drivers in certain regions. Although they come in large quantities, Internet-based data tend to have many selectivity, security and data privacy concerns, which make it difficult to use this type of data for official statistics purposes. Mobile phone-based data make for an abundant and pervasive source of data on mobility because they can provide the

[1] "A population is 'hidden' when no sampling frame exists and public acknowledgement of membership in the population is potentially threatening" (Heckathorn, 1997:174).

closest picture of reality with near real-time data and relatively limited selectivity issues, given the widespread use of mobile phones around the world. Thus, for migration statistics, mobile phones or MPD probably constitute the most useful data source.

Mobile positioning data

General description of mobile positioning data

Mobile phones generate various data traces that can be analysed in an active or passive way – making or receiving a phone call, sending or receiving an SMS, connecting to the Internet, connecting to a different antenna, turning the phone on, technical communication between a mobile device and a mobile network, etc. Of most interest to the analysis of migration and mobility are the data that mobile network operators (MNOs) collect about the locations of mobile network subscribers for operational purposes: MPD.

The most common types of MPD used are CDRs and signalling data. We introduce these below, and more detail about them can be found in the first version of the United Nations Global Working Group handbook (2019).

CDRs are widely used in research, since all MNOs store records of calls, text messages and data sessions for billing purposes. Oftentimes, national data-retention policies request MNOs to keep such data for a year or more.

Signalling data are generated as a result of the continuous communication between a mobile device and a mobile network to maximize signal quality. The records created are similar to CDRs. However, signalling data provide higher density data than CDRs because one data point is generated every few seconds, or minutes, when a mobile phone is switched on and is within a signal coverage area.

Sociodemographic data may include some basic descriptive characteristics like age, sex, preferred language and often some usage classification (contract type, services used, average invoice, etc.). Practically, these indicators are available only for domestic (and thus also outbound) subscribers, but they can often be misleading. For example, if the head of the household signs a contract for mobile services for the whole family, all family members might be characterized as the head of the household (e.g. typically middle-aged male). Further, for inbound subscribers, MNOs do not have such information, except for the home country of the SIM card user.

From the point of view of an MNO in a specific country, MPD come in three forms (Figure 1):

(a) Domestic data: MPD of the subscribers of the home network of the MNO within the country limits.
(b) Outbound roaming data: MPD of the domestic subscribers using a roaming service of other, mostly foreign, MNOs.
(c) Inbound roaming data: MPD of subscribers from other, mostly foreign, MNOs.

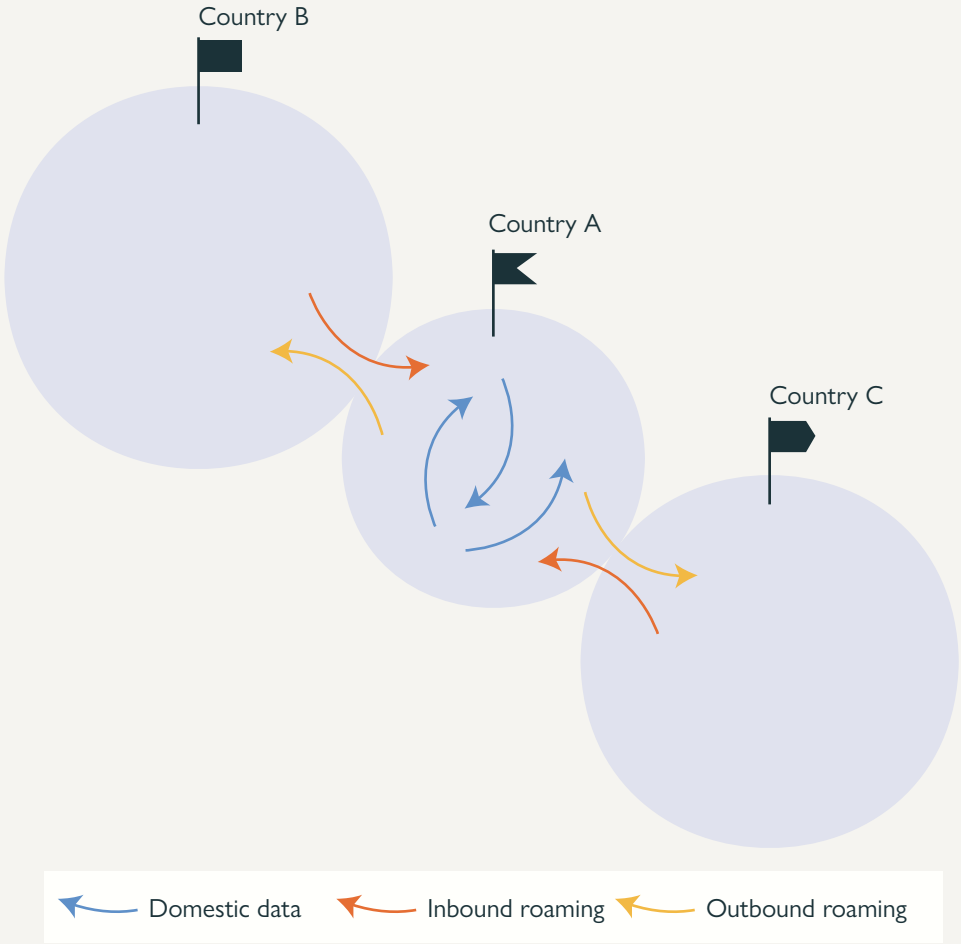
Figure 1. Different forms of data from the mobile network operator of Country A

Mobile positioning data (MPD) come in three forms:

DOMESTIC DATA
MPD of the subscribers of the home network of mobile network operators (MNOs) within the country limits.

INBOUND ROAMING DATA
MPD of subscribers from other, mostly foreign, MNOs.

OUTBOUND ROAMING DATA
MPD of the domestic subscribers using a roaming service of other, mostly foreign, MNOs.



It is important to note that an MNO has data about its subscribers in the domestic network within the national borders (domestic data) and roaming in other mobile networks (outbound roaming data). The MNO will also have data about subscribers from other providers roaming in their network (i.e. inbound roaming data). MNOs store domestic, inbound and outbound data available as CDRs. MNOs might have signalling data, but this is not guaranteed. This means that to a certain extent, MPD can traverse borders and are suitable for the analysis of international mobility. The availability of the data is presented in Table 1.

Table 1. Description of availability of various data forms and sources

	Domestic	Outbound	Inbound	Additional notes
Subscriber activity data (call detail records)	• Always available	• Always • Are sent to the operator	• Always • CDRs that operator sends to another MNO	Good for long-term statistics; not good for short term
Network data (signalling)	• Usually available • Depends if MNOs have implemented signalling	• Mostly not • Available if roaming partner MNO has implemented signalling and is providing the data for its roaming partners	• Usually • If available for domestic, then also available for inbound	Good for short- and long-term statistics
Sociodemographic data (client relationship management data)	• Yes, but unreliable • Customer information usually is not thoroughly checked for quality and representativeness		• Not available	Sociodemographic data are often unreliable

The location is almost always given by the position of the mobile network cell that the subscriber is connected to. Depending on the density of the data, there are always gaps in time between records. Deriving any kind of relevant information from this initial data set requires conceptual modelling and an understanding of what these data represent, as well as a transformation of the initial MPD into a data model that corresponds with this conceptual understanding.

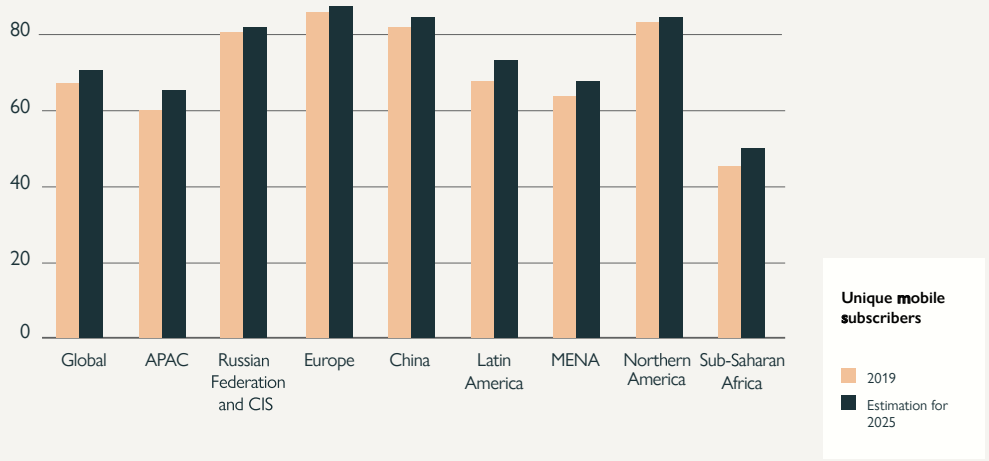
Possibilities and limitations of using mobile positioning data

Possibilities

MPD are collected from a large sample. MPD arguably represent the best data source for understanding people's movements in space and time mainly because they cover the majority of the population (GSMA, 2020), with a large amount of information representing the whereabouts of people. The population census, which is a key source of migration data for most countries, has complete coverage of the population but is conducted every 5 to 10 years. No other data source – including administrative registers, surveys or novel data sources – offers such a large sample representation.

Figure 2. Unique mobile subscribers for 2019 and estimation for 2025

MPD are the best data source for understanding people's movements in space and time mainly because they cover the majority of the population (GSMA, 2020) with a large amount of information representing the whereabouts of people.



Source: GSMA, 2020.

MPD are available globally, and most countries in the world have very high mobile phone penetration rates (Figure 2). The standards for storing CDRs are similar because of global standards. This means that methods for data analysis that work in one country can easily be applied in another.

MPD can be analysed to study several different mobility phenomena at the statistical level. This means that a good data model based on MPD can provide insights on various current data gaps – one-off or regular mobility, national or international movements, short- or long-term migration.

MPD can be relatively safe in terms of user privacy. Although there are obvious privacy concerns when CDR data are shared for research and other purposes, appropriate safeguards can be taken, such as pseudonymization, masking and data privacy impact assessments. These have so far ensured that there have not been major privacy breaches connected to research and statistical work with MPD.

MPD are a time machine. Since this source relies on historical data collected passively over a long period, it enables both near real-time analyses as well as going back as far as the database allows. This means that any phenomena happening currently can be compared to a benchmark period. It also means that methods can be refined and results recalculated at a later point in time.

Limitations

MPD also have several limitations.

MPD do not hold qualitative data about the subscribers' motivations and intentions, and generally provide limited demographic information. MPD truly are a quantitative data source that can be mixed with qualitative methods.

Statistically, MPD still present selectivity issues. There is inherent selectivity in mobile phone usage, and each MNO has a different subscriber base in terms of sociodemographics and geography. In fact, selectivity issues can further be exacerbated through the analysis – for example, by relying too heavily on records about data usage, skewing the results towards more affluent smartphone users.

MPD are most reliable within a country. When people cross borders, they tend to switch their SIM cards to use a local one, especially with long-term journeys or migration. Although this is not always the case and there are exceptions, usually when people migrate to another country permanently (or for an unforeseen period of time), they obtain a local SIM card and discontinue their home SIM card. These subscribers will not appear in the outbound roaming data of their original home country MNO; they will appear only in the domestic (and not inbound roaming) data of the new country's MNO, making it impossible to track migration using MPD. However, many migrants who move to a new country for a limited time (regardless if short or long term) do keep their original SIM cards and may therefore be identified in migration indicators from MPD.

MPD can be difficult to access. The most important limitation is connected to privacy and other legal considerations, as MPD include highly sensitive information about the whereabouts of the subscribers, which MNOs can be reluctant to share with researchers.

MPD include noise and erroneous records. Because they are not collected for statistical purposes, special care should be taken to clean the data of errors and noise, which might affect the final result.

Due to such limitations, MPD should be analysed with caution. They should ideally be viewed as a complementary source to traditional data sources, although they can be used independently if their limitations are accounted for. That is why analysing MPD requires a combination of technical and statistical knowledge.

Methodology for using mobile positioning data

The statistical community adheres to the Fundamental Principles of Official Statistics (United Nations, 2014, 2015), while MNOs often optimize for the risk in sharing their data. There is a fine balance between statistics and privacy that is often the concern in MPD projects.

Work under way at the United Nations Global Working Group on Big Data for Official Statistics determined that to use MPD for any kind of statistical indicators, the following principles should be kept in mind:

(a) **Necessity and proportionality**: Making sure MPD are relevant and fit for purpose to provide the necessary timely, frequent and geographically detailed information for the monitoring of human mobility.
(b) **Professional independence**: Guaranteeing the transparency of methods and practices between statistical or data entities.
(c) **Privacy protection**: Preprocessing and aggregating data to protect privacy, technical and governance frameworks that ensure confidentiality and data security, and presenting data in a manner that avoids the possibility of individuals being reidentified.
(d) **Commitment to quality**: Acknowledging and communicating the limitations of MPD, and planning for future improvements in methodology when time and resources become available.
(e) **International comparability**: Being transparent about methodology and maintaining the reproductivity of results and/or following internationally recognized guidelines for MPD for official statistics.

There are broadly two main approaches to processing MPD:

(a) Using aggregated data from MNOs to produce limited indicators that only describe some basic concepts. For example, extracting information on basic movement patterns between countries or creating a simple mobility index as has been done in several countries during COVID-19. This requires a simple set-up, demands fewer resources and avoids many privacy concerns.
(b) Using pseudonymized raw MPD with microlevel data, with extensive methodology and a data model reflective of the reality, which provides a possibility to acquire a variety of indicators, classifications and aggregation options. This kind of set-up enables more research-driven applications of MPD. Although it is more resource-intensive in the short term, it creates opportunities to have a solid method in the long term.

When raw MPD are used, many methodological approaches can be adopted for processing the data. Ideally, the methodology results in a universal, reality-reflecting data model that can be used to produce a variety of indicators for different domains. But the methodology can also be simpler and focus on producing the results for a specific domain and indicators.

Specifically for migration, the following methodological considerations are required to be able to produce insightful migration indicators:

(a) The period of MPD must be reasonably long to be able to make any assumptions. It is impossible to measure migration if MPD cover a very short period (e.g. a week). The longer the period, the more reliable the assumptions made from MPD.
(b) Identification of the subscribers' country of residence, and the period for which the residency is valid. We can take the definition of residency[2] and adjust it to MPD, meaning a person who spends the majority of a 12-month period in a territory is considered a resident of said territory.

[2] "Being present for one year or more in a territory or intending to do so is sufficient to qualify as having a principal dwelling there." (IMF, 2009)

(c) Identification of place of residence (home) and other important locations like work and school, other regularly visited places (a.k.a. anchor points/meaningful locations), and the periods for which these locations are valid.
(d) Identification of the usual environment where the subscriber dwells.
(e) A data model including stays, indicating a presence in a country and specific locations for a period of time.
(f) Bias identification algorithms that eliminate MPD records or other data model objects (subscribers, trips, stays, etc.) that are technically or logically erroneous (e.g. accidental roaming near borders, seamen using roaming services offshore).
(g) Application of migration concepts, definitions, parameters, etc., that data model objects (subscribers, trips, stays, etc.) correspond to (international migration is when a country of residency is changed, transnational subscriber is when the subscriber spends a certain amount of time in and travels a specific number of times between at least two countries, etc.).

A good practice is that the methodology used for processing MPD and the data model can be employed for several domains apart from just a single research question on migration. Data models can be the following:

(a) Domain-specific for a single domain: MPD are processed specifically for one purpose, and they are not usable or could be deceptive when used in other domains.
(b) General: MPD are processed holistically, and the data model takes into account the requirements, concepts and necessary processes of all domains where MPD can be used.

Mobile positioning data for multiple domains in Indonesia

The Indonesian national statistical office, Badan Pusat Statistik (BPS – Statistics Indonesia), uses MPD for several domains. Underlying the use is a long-term partnership with the MNO, an international expert in analysing MPD, and a data model built on existing research in the field. The partnership enables BPS to develop statistics for several domains – information and communications technology statistics, population, commuting, tourism statistics and future migration – through a single data model. According to BPS, Indonesia's case is also a telling tale about the necessity to have in place a sound methodological footing that corresponds to the conceptual frameworks of domain statistics. MPD are a daily production of telecommunications companies, the MNOs, so a clear, trusted and transparent methodological framework is required. Quality assurance and data validation must be based on and derived from the conceptual framework adopted and implemented.

Using mobile positioning data for migration statistics

From an MPD perspective, the concept of migration can be defined as short or long term, and it can be represented by national or international mobility of the subscribers (Table 2).

Table 2. Comparison matrix on the readiness of using mobile positioning data for short term versus long term and national versus international migration

	National	International
Short-term mobility	**Domestic commuting, domestic tourism/ seasonal migration** Existing studies on the identification of important locations (homes) with good methodology. CDRs are not so helpful to analyse short-term mobility. Signalling data can be more useful for that. MPD are relatively well researched and shown to help determine migration stocks and flows. Absolute numbers require good estimation methods; relative numbers can be sufficient.	**Cross-border commuting, international tourism/seasonal migration, transnationalism** MPD are somewhat researched and provide an indication of migration flows, but it is difficult to determine stock due to undercoverage issues.
Long-term mobility	**Internal migration** Several studies on the identification of important locations (homes) with good methodology. CDRs and signalling data can be used for tracking changes of residence over time. Absolute numbers may not be reliable; relative numbers can be sufficient.	**Emigration/immigration/migrant workers** MPD are not well researched, and there are data source limitations in detecting migration due to severe undercoverage.

National migration

State of the art

MPD have been used in several academic studies and practical applications that focus on the identification of regularly visited locations (important, meaningful locations and anchor points, i.e. home, work), mobility between those locations (commuting, short-term migration), and change of those locations over time. One of the main advantages is that MPD offer a statistically accurate representation of the distribution of people in an area and can be used to track large and heterogeneous groups of people (Bianchi et al., 2016). However, MPD do not represent the general population, thus they are most useful for identifying relative numbers in migration, with the possibility of extrapolating the general population (absolute numbers) using some reference (ground truth data) for calibration.

Case studies

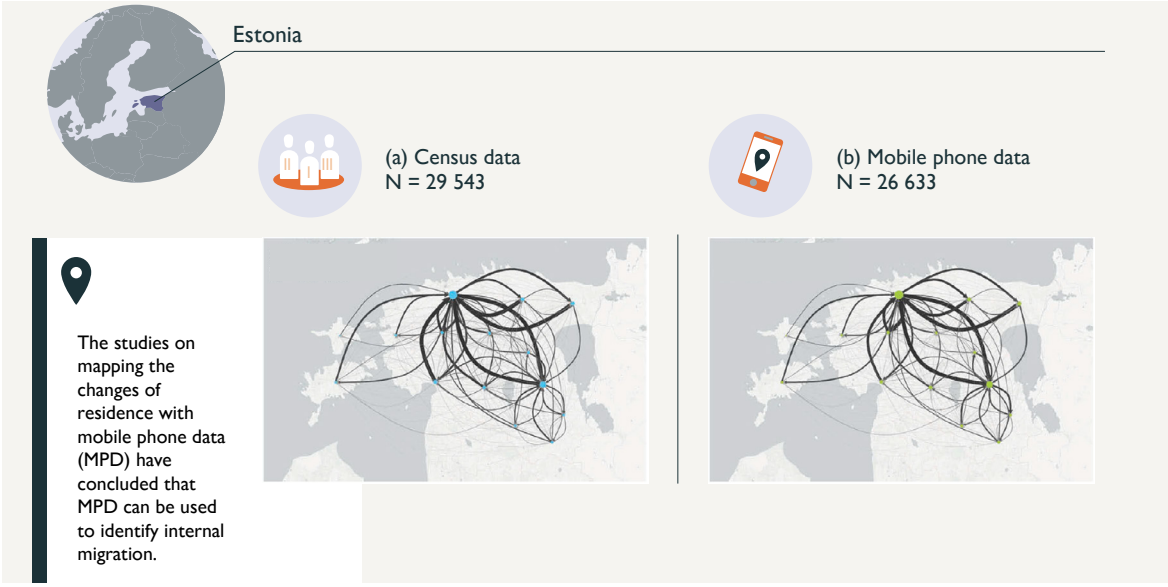
(a) There have been several successful studies on the identification of important and meaningful locations using MPD:

(i) Several methods have been proposed and tested to identify the meaningful locations for subscribers using MPD (Ahas et al., 2010; Bianchi et al., 2016; Mamei et al., 2016; Jiang et al., 2017) (see Figure 3).

(ii) Identification of home locations using MPD correlates to the number of residents in administrative units compared to population census (Ahas et al., 2009).

(b) The studies on mapping the changes of residence with MPD have concluded that MPD can be used to identify internal migration (Kamenjuk et al., 2017) (see Figure 3).

Figure 3. Comparison of flows for internal migration between counties in Estonia according to the 2011 census and mobile positioning data during the period 2010–2011



Note: N = number of migrants between counties (Local Administrative Unit Level 1) (Kamenjuk et al., 2017:15).

These maps are for illustration purposes only. The boundaries and names shown and the designations used on these maps do not imply official endorsement or acceptance by the International Organization for Migration.

(c) A study based on data from the United States of America and Senegal takes another approach, suggesting that short-term flows may be useful in modelling long-term flows (Fiorio et al., 2020).

(d) Use of MPD for national migration by the University of Southampton and Flowminder in Namibia found that estimates of relative (not absolute) internal migration flows based on CDR-derived models matched well with relative migration flows from census data and showed the potential for MPD to update relative intercensal migration numbers. The latter would need separate validation (Lai et al., 2019).

What can potentially be measured

(a) Estimation of the de facto population;

(b) Estimation of residents;

(c) Estimation of commuters between home–work and home–second home;

(d) Estimation of short-term changes in residency (e.g. seasonal migration or displacement) beyond administrative borders per time period (month), their trips and duration of stay;

(e) Number of long-term (permanent) changes in places of residence beyond administrative borders per time period (month, quarter, year).

Methodological challenges

The main methodological challenges are connected to identifying important locations, most significantly the place of residence (home), and the identification of the time period when a certain place is home and when it has changed. This has been rather successful, although the classification of those important locations can be a challenge in itself.

Another challenge is related to extrapolation to the general population. Data from MNOs do not fully represent the general population, and statistical extrapolation methods should be developed to derive absolute numbers. Even with good extrapolation methods, in the absence of reliable reference data to calibrate MPD results, relative numbers regarding migration (or any other phenomena) are more reliable (e.g. the number of subscribers with changed home location compared to the total number of home locations in an area).

Recommendations

(a) The identification of important locations (homes) of subscribers has proven to be successful.
(b) Short-term mobility can be identified relatively accurately when signalling data are used.
(c) Change of important locations in time (i.e. long-term migration) can be identified using CDRs and signalling data.
(d) Extrapolation to the general population is problematic. The limitations of MPD and the reliability of absolute numbers provided by MPD results should be considered. Relative numbers might be used when no reliable reference data can be utilized for calibration.

International migration

State of the art

International migration has been measured using MPD in several cases. However, MPD are mostly helpful to estimate short-term (temporary) migration, rather than long-term (permanent). The reason is that people tend to acquire local SIM cards when permanently migrating to different countries, so the consistency of the subscriber's identity is lost in the data set. Nevertheless, there are options for partially measuring permanent migration and providing insights or trends that would otherwise be unavailable, especially in low- and middle-income countries.

Case studies

(a) Measuring transnational migration using MPD from Estonia to neighbouring countries (University of Tartu and Positium). The study shows the possibility of classifying "transnational" individuals based on the number of trips taken and time spent abroad (Ahas et al., 2018). Mooses et al. (2020) went on to distinguish the transnational activity space of different groups based on the language they speak, an ethnographic parameter available to the researchers.
(b) Research on temporary cross-border mobility (migration) between Estonia and Finland where travellers were classified as tourists, commuters, transnationals and long-term visitors (Silm et al., 2021).
(c) Mobility data are particularly useful to study groups that are not easily covered in the national surveys, such as refugees, as in the international study of Syrian refugees living in Türkiye that had been labelled by the MNO Turkcell based on tariff or registration (Salah, n.d.).

What can potentially be measured

(a) Mobility across the border, classified into tourists, commuters, transnationals and long-term visitors.
(b) Cross-border commuters who live in one country and work in another country, their trips, and duration of stays in each country.
(c) Trend of long-term (permanent) changes in countries of residence.

Methodological challenges

The main challenge of using MPD to measure international migration is related to the quality of MPD for long-term international travels, specifically the use of local SIM cards instead of preserving home MNO SIM cards, which introduces inconsistencies in the subscriber's time series.

Extrapolation to the general population is also a challenge, as the number of mobile devices or SIM cards is never equal to the number of people.

Recommendations

Short-term international mobility (e.g. cross-border commuting, seasonal migration) can be analysed through the use of MPD. However, the limitations of MPD for this subject should be considered.

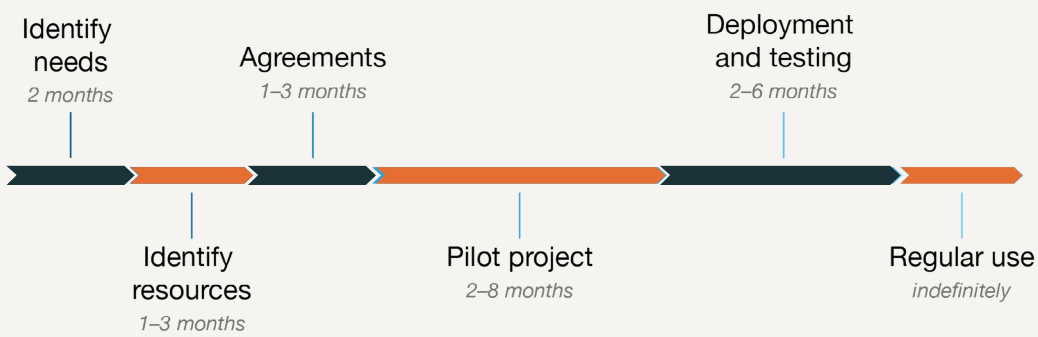
Long-term international migration has largely not been investigated using MPD because of the limitations of MPD in this area. Still, in the absence of other data, MPD can provide some useful insights. For recurring migration (transnationalism), MPD are considered as a good source, mainly because people regularly dwelling across several countries tend to preserve their SIM cards.

Some assumption-based methods might also be employed here. For example, the disappearance of foreign SIM cards or the appearance of new domestic SIM cards can suggest the arrival of migrants in specific areas of a country.

Road map for the use of mobile positioning data for migration

Due to the nature of big data, dealing with all the possibilities and limitations of these data in one research project can be overwhelming. That is why we recommend an analytical approach that integrates lessons from previous work to build a long-lasting data pipeline (Figure 4).

Figure 4. Indicative timeline of a road map for implementation of the mobile positioning data processing system



A general road map for implementing MPD could involve the following steps. The given timeline is indicative and largely depends on the legislation, readiness of data providers to cooperate, funding and other circumstances:

(a) Identify needs (two months). This should focus on identifying (i) short-term needs for proof of concept and low-hanging fruits that can be achieved quickly; and (ii) long-term needs for data on selected migration indicators, while also adopting a broader view, crucial for building a single data model for several domains and realizing the full value of MPD.

(b) Identify key resources and constraints (one to three months):
   (i) Identification of key players (MNOs as data providers, possibly national statistical offices (NSOs), telecommunication regulators, academia, etc.).
   (ii) Legislation analysis: Does national legislation regulate and allow for the use of MPD for statistical purposes, and if so, what are the constraints?
   (iii) Identification of human resources:
      a. Technical skills to implement and operate the pilot and the production system.
      b. Analytical skills to analyse and derive meaningful information from the data set.
   (iv) Identification of computational resources available to process such amounts of data within a reasonable period. Where does the processing take place (physically)?
   (v) Identification of the fundamental principles for statistical production and how to ensure adherence to those principles.

(c) Agreements between key stakeholders for conducting a pilot project (one to three months). These include MNOs and possible partners involved in the processing and/or analysis of the data. This phase involves defining roles and responsibilities, success criteria for the pilot project, etc.

(d) Conduct a pilot for proof of concept (two to eight months). This includes analysis of the pilot project results and of the feasibility for continuous production; determining the funding of the continuous data processing system (deployment, implementation, upkeep); and agreements with data providers for regular data feed, and with other involved parties. A pilot project should include data from several years to be able to historically compare the time series, seasonality, etc.

(e) Deployment and testing of the continuous data processing system (two to six months).

(f) Regular use of the continuous data processing system and implementation (indefinitely). MPD data are regularly provided by MNOs (on a monthly basis or another frequency), and the processing is conducted in a central (data from MNOs are collected and stored in a single physical server) or distributed (within the premises of each MNO) processing system.

Time is needed to define the best arrangements with the MNOs regarding the trade-off between privacy protection and mobile operator data requirements. Full open access to the data would give the NSO and its service provider all possibilities to develop detailed and relevant indicators. Fully anonymized and aggregated data limit the possibilities to develop the desired indicators. The restrictions imposed on the data to protect privacy may affect the quality of the indicators, which can be compiled for policy purposes. Therefore, a good solution between protecting privacy and

guaranteeing the quality of the indicators should be worked out. Privacy-preserving techniques can help to safeguard privacy while allowing more elaborate operations on detailed mobile operator data. Moreover, well-established legal frameworks for data protection in a country can be helpful in devising a good arrangement between MNOs and NSOs.

The presented road map does not take into account unforeseeable obstacles that almost always occur and may stall the process for months or even years, such as the need to resolve legal issues at the government level, or technical (or administrative) preparations of MNOs to provide the necessary data (e.g. signalling data are not yet available, but MNOs are implementing these, and it will take time).

Over the years, several aspects of national and international migration have been studied, and methods have been developed showing the usefulness of mobile phone data to analyse migration and mobility patterns. More research is needed specifically in some areas – e.g. the study of international migration, models for estimating the total population, migration in low-income country contexts. Setting up continuous data pipelines allows for such experimentation and research into necessary topics, while benefiting from the value of existing applications in domains where adequate methods have already been developed.

While no data set is perfect, mobile phone data can already be used in most countries to provide a timely, factual assessment of national and international movements. For that, agreements need to be made and data pipelines set up to provide quality statistics in the long term.

REFERENCES*

Ahas, R., S. Silm, O. Järv, E. Saluveer and M. Tiru

2010 Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17(1):3–27.

Ahas, R., S. Silm, E. Saluveer and O. Järv

2009 Modelling home and work locations of populations using passive mobile positioning data. In: *Location Based Services and TeleCartography II: From Sensor Fusion to Context Models* (Gartner, G. and K. Rehrl, eds.). Springer, Berlin, pp. 301–315. Available at https://link.springer.com/chapter/10.1007/978-3-540-87393-8_18.

Ahas, R., S. Silm and M. Tiru

2018 Measuring transnational migration with roaming datasets. *Adjunct Proceedings of the 14th International Conference on Location Based Services*. ETH Zurich, pp. 105–108. Available at www.research-collection.ethz.ch/handle/20.500.11850/225599.

Bianchi, F.M., A. Rizzi, A. Sadeghian and C. Moiso

2016 Identifying user habits through data mining on call data records. *Engineering Applications of Artificial Intelligence*, 54:49–61.

Fiorio, L., E. Zagheni, G.L. Abel, J. Hill, G. Pestre, E. Letouzé and J. Cai

2020 Analyzing the effect of time in migration measurement using geo-referenced digital trace data. MPIDR Working Paper WP 2020-024. Max Planck Institute for Demographic Research, Rostock (forthcoming in *Demography*). Available at www.demogr.mpg.de/papers/working/wp-2020-024.pdf.

Global Migration Group

2017 *Handbook for Improving the Production and Use of Migration Data for Development*. Global Knowledge Partnership for Migration and Development (KNOMAD), World Bank, Washington D.C. Available at www.knomad.org/publication/handbook-improving-production-and-use-migration-data-development-0.

GSMA

2020 *The Mobile Economy*. Available at www.gsma.com/mobileeconomy/.

Heckathorn, D.D.

1997 Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):174–199.

** All hyperlinks were working at the time of writing this report.*

International Monetary Fund (IMF) (ed.)
2009 *Balance of Payments and International Investment Position Manual*. Sixth edition. Washington, D.C.

Jiang, S., J. Ferreira and M.C. Gonzalez
2017 Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data*, 3(2):208–219.

Kamenjuk, P., A. Aasa and J. Sellin
2017 Mapping changes of residence with passive mobile positioning data: The case of Estonia. *International Journal of Geographical Information Science*, 31(7):1425–1447.

Lai, S., E. zu Erbach-Schoenberg, C. Pezzulo, N.W. Ruktanonchai, A. Sorichetta, J. Steele, T. Li, C.A. Dooley and A.J. Tatem
2019 Exploring the use of mobile phone data for national migration statistics. *Palgrave Communications*, 5:34.

Mamei, M., M. Colonna and M. Galassi
2016 Automatic identification of relevant places from cellular network data. *Pervasive and Mobile Computing*, 31:147–158.

Mooses, V., S. Silm, T. Tammaru and E. Saluveer
2020 An ethno-linguistic dimension in transnational activity space measured with mobile phone data. *Humanities and Social Sciences Communications*, 7(1):140.

Salah, A.A.
n.d. Mobile data challenges for human mobility analysis and humanitarian response. In: *Handbook of Migration and Technology* (McAuliffe, M., ed.). Edward Elgar Publishing (forthcoming).

Silm, S., J.S. Jauhiainen, J. Raun and M. Tiru
2021 Temporary population mobilities between Estonia and Finland based on mobile phone data and the emergence of a cross-border region. *European Planning Studies*, 29(4):699–719.

United Nations
2014 Fundamental Principles of Official Statistics. A/RES/68/261. Available at https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf.
2015 *United Nations Fundamental Principles of Official Statistics: Implementation Guidelines*. Available at https://unstats.un.org/unsd/dnss/gp/Implementation_Guidelines_FINAL_without_edit.pdf.

United Nations Global Working Group on Big Data for Official Statistics
2019 *Handbook on the Use of Mobile Phone Data for Official Statistics*. Available at https://repository.unescap.org/handle/20.500.12870/4236.

AUGMENTING MIGRATION STATISTICS USING SOCIAL MEDIA

Jisu Kim,[1] Emilio Zagheni[2] and Ingmar Weber[3]

Introduction

Migration is a key driver of demographic and social change, as well as an often hotly debated political issue. Over the past years, the phenomenon has attracted attention from policymakers, researchers, as well as the general public, particularly due to increasing numbers of forcibly displaced individuals. At the end of 2019, about 272 million people migrated (DESA, 2019a), and about 79.5 million people were forcibly displaced from their homes, either internally or internationally, due to conflicts, generalized violence or human rights violations.[4] These numbers have continued to increase (DESA, 2019b), up until the outbreak of COVID-19 which seems to have halted this growth, although there is speculation that we may observe a spike in the number of migrants once the travel restrictions are eased (O'Brien and Eger, 2020). In order to cope with the complexity of the matter, the need for up-to-date and rich data to better monitor and manage the situation has become clear. However, traditional migration data such as census, register and survey data that researchers mainly rely on have a number of limitations and imperfections (Sîrbu et al., 2020).

First, the definitions used may be inconsistent across different countries. For example, the reported number of people immigrating to Spain from Italy in 2015 was 17,350 according to official data from the Spanish National Statistics Institute. The same number was estimated to be only 5,003 according to official data from the Italian National Institute of Statistics.[5] This discrepancy is in part related to the varying definitions of "migrant" and the different ways of collecting data. It is also due to the fact that there are limited incentives for emigrants to declare their departure in their origin country, and more incentives for immigrants to register in the country of destination. Apart from inconsistencies, traditional data are often time-consuming, costly to collect, and published with delay. In extreme circumstances, there are no data at all.

[1] Jisu Kim is a research scientist at Max Planck Institute for Demographic Research. Her research focuses on the intersection of migration sciences, the economics of migration, complex social networks, statistical models and data-driven algorithms. Jisu holds a PhD in Data Science from the Scuola Normale Superiore in Italy.

[2] Emilio Zagheni is Director of the Max Planck Institute for Demographic Research in Rostock, Germany, and Head of the Laboratory of Digital and Computational Demography. He is best known for his pioneering work on using web and social media data for studying migration processes and for his role in developing the field of digital and computational demography.

[3] Ingmar Weber is an Alexander von Humboldt Professor for Artificial Intelligence at Saarland University. Previously he was Research Director for Social Computing at the Qatar Computing Research Institute. He studied and worked at the University of Cambridge, the Max Planck Institute for Informatics, the Ecole Polytechnique Fédérale de Lausanne and Yahoo Research.

[4] More information is available at www.unhcr.org/globaltrends2019/.

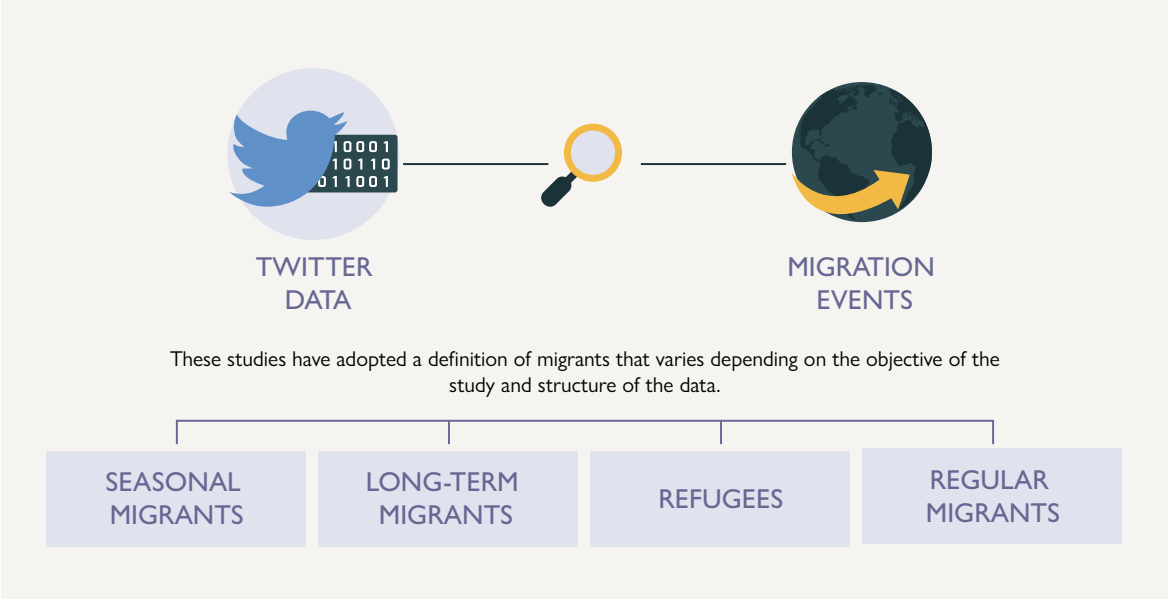
[5] More information is available at https://ec.europa.eu/eurostat/home.

As we live in the digital age, a vast amount of passively collected data exists from new sources, such as call detail records, social media platforms and other Internet services. Some of them are freely available and enable us to gather real-time data. Furthermore, these sources provide large-scale data on a granular level, which enables scientists to study phenomena that are hard to analyse using more coarse-grained data. Aside from potential advantages such as the low cost, temporal and spatial granularity, and real-time availability, these non-traditional data sources come with a range of challenges, particularly in relation to selection biases and ethical issues.

This chapter focuses on the potential value of data from two social media platforms: (a) geotagged tweets and (b) Facebook ads audience estimates – to demonstrate how these data can be used for migration statistics. The declared goal here is to enhance, not replace, traditional migration statistics. The chapter is divided into two main sections; the first section focuses on Twitter and geotagged tweets, whereas the second section concentrates on Facebook ads audience estimates. Each section discusses how the data can be accessed, provides details on the drawbacks of the data, and describes how such data have been used to produce estimates and enhance migration statistics. We conclude the chapter with a summary of overall pros and cons of social media data and the potential of these data for exploring additional dimensions of migration processes that are difficult to quantify with traditional sources only.

Accessing and using Twitter data to measure migration

In the past years, several studies have used Twitter data to study migration events (Fiorio et al., 2017; Hausmann et al., 2018; Hawelka et al., 2014; Huang et al., 2020; Lamanna et al., 2018; Kim et al., 2020; Mazzoli et al., 2020; Zagheni et al., 2014). These studies have adopted a definition of "migrant" that varies depending on the objective of the study and structure of the data, and have focused on various groups of migrants, such as seasonal migrants, long-term migrants, refugees or regular migrants. Studies commonly try to pinpoint a country of residence by looking at the most frequent location of tweets for a period considered. For instance, Zagheni et al. (2014), Hawelka et al. (2014), and Kim et al. (2020) all define "country of residence" as the country where a user tweets for the most part over a period of four months (Zagheni et al., 2014) or a year (Kim et al., 2020). According to Zagheni et al. (2014), outmigration events are detected when there is a change in the country of residence.



On the other hand, migration events are detected, according to Kim et al. (2020), when the nationality identified is different from the country of residence – where "nationality" is defined by "linguistic and social connections to a migrant's country of origin". As reported by Hawelka et al. (2014), a mobility pattern is detected once a country other than the country of residence appears in tweets. However, different from the other works, Hawelka et al.'s article (2014) studied global mobility and was able to detect seasonal mobility patterns which fall under the definition of "visitors" rather than migrants, under the definition set by DESA's *Recommendations on Statistics of International Migration* (1998). Similarly, Mazzoli et al. (2020) have observed routes of migrants and refugees during the migration crisis in the Bolivarian Republic of Venezuela by detecting a change in location during the time window of observation. The emigration identification strategy is somewhat similar also in Hausmann et al.'s working paper (2018). Although they were not focusing on migrants, strictly speaking, Huang et al. (2020) showed that Twitter data can be a good data source to study how mobility patterns have changed after international efforts to reduce mobility were implemented during the middle of the COVID-19 pandemic.

A challenge common to all social media-based studies, in particular those using Twitter data, is how to validate the results. In other words, how can one make sure that the results are not biased and that, ideally, the findings can be generalized to the wider population, not only a small subset of Twitter users? As the studies mentioned above have shown, there are numerous ways to deal with biases. For instance, Zagheni et al. (2014) employed the difference-in-differences technique to control for different Twitter usage across countries by assuming that, in the short term, differences in bias across countries may be constant. Other studies have validated their results by comparing the predicted results with official statistics (Hawelka et al., 2014; Kim et al., 2020; Mazzoli et al., 2020). In a recent contribution, Hsiao et al. (2020) showed that once spatial and temporal bias structures are statistically modelled from Twitter data, other data sources such as survey data can be combined to develop a joint model. In addition, the authors concluded that not only did the joint model correct the upward bias in Twitter data, but it also outperformed the accuracy level of survey data alone.

In terms of data collection, Twitter is a freely available data source that can be accessed using an application programming interface (API).[6] The two main methods of accessing data using the Twitter API are through the search API and the streaming API. Both APIs return data in JSON format, called "objects",[7] which are easy to store and manipulate.

The search API enables the collection of publicly available tweets and profiles of users. The search can be done on a specific user through either the user ID or the username. This returns a user object that contains information about an individual's profile, such as when the account was created; the number of tweets, followers and friends; as well as the location that the user has declared to be in. Otherwise, the search can be run on specific keywords or geolocations embedded in tweets. The geolocations can be specified – a country, place, bounding box, or within a radius of 0.01 km up to about 40 km from specific coordinates.[8]

[6] More information is available at https://developer.twitter.com/en/docs/twitter-api.

[7] More information is available at www.json.org/json-en.html.

[8] More information is available at https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location.

On the other hand, the streaming API allows us to gather random samples from 1 per cent of all new public tweets in near real time. The streaming API also allows us to specify filter criteria (e.g. keywords, geolocations, and user IDs or usernames). However, unlike the search API, it returns tweets matching the filter criteria as soon as matching tweets are created. The returned content of matching tweet objects includes the tweet text, location information (where present), the language in which the tweet was written when it was created, and additional information, such as whether the tweet was part of a thread. It also contains the entity object which lists tweet details such as hashtags, URLs and mentioned IDs. To collect small amounts of data, instead of using the APIs, there are websites to search for tweets (e.g. on Twitter directly),[9] or where detailed searches for particular users can be issued (e.g. on followerwonk).[10]

Figure 1. An example of a search on followerwonk.com for "IOM GMDAC"

After having obtained the data, there are few notes of caution to consider. First, only a small percentage of tweets come with geolocations based on the user opting in to share their exact position. For instance, Morstatter and Liu (2017) showed that only about 3.2 per cent of tweets from the streaming API are geotagged. This means that any given user is unlikely to have geotagged tweets and, correspondingly, challenges related to self-selection bias need to be addressed.

Second, it requires effort to clean and process the data. Often, the tweets are not directly usable as they are "noisy" and/or incomplete. For instance, tweets contain repeated characters (e.g. "wooooooow"), typos, or Internet slangs that are not familiar to everyone, and these pose challenges to standard natural language processing (NLP) tools. Some tweets may also be incomplete in that they require additional context, such as a thread of tweets, to make sense of them. Cleaning and removing data may result in a considerable loss of information. Further, bots or spam accounts introduce additional data quality issues. It is also important to make sure that identifying migration events is not confounded with misleading activities.

Another limitation is that Twitter does not provide user attributes such as education or income level, which are often helpful for more in-depth migration studies. Nevertheless, certain characteristics, such as age, ethnicity or sex, can often be inferred with reasonable accuracy using the profile image (Zagheni et al., 2014; Huang et al., 2014).

[9] More information is available at https://twitter.com/explore.

[10] More information is available at https://followerwonk.com.

Lastly, there are privacy issues. It is vital to make sure that no personal information obtained from the data is published even if Twitter data are openly available. A proper infrastructure is needed where data can be safely stored in a secured server. Additionally, to comply with Twitter's Terms of Service, only "dehydrated" data sets can be shared for research or archival purposes, except for small collections not exceeding 50,000 tweets.[11] This requires data to be in the form of unique IDs which can then be "rehydrated" – in other words, restored to the original data.[12] This gives the user a chance to opt out of subsequent studies by deleting their tweet/account. It should also be noted that while most Twitter content is public and accessible to anybody, an individual user might not expect researchers to algorithmically collect and analyse their tweets. How to best address these expectations of data use, which are separate from legal considerations, remains a challenge, with answers depending on the specific context.

Accessing and using Facebook ads audience estimates to measure migration

Given the size of Facebook, with more than 2.7 billion users in September 2020,[13] researchers have also tried to tap into data from the social networking site, in particular through its advertising platform. Here existing works have shown that Facebook can indeed be a good alternative data source to study migration phenomena (Alexander et al., 2019; Gendronneau et al., 2019; Palotti et al., 2020; Spyratos et al., 2018, 2019; Zagheni et al., 2017). The studies have all shown that stocks or flows of migrants identified through Facebook data are indeed well correlated with official statistics in various cases. For instance, Zagheni et al. (2017) found that 94 per cent of the variability in data from the American Community Survey (ACS) is explained by Facebook's estimates. In other words, the relative trends observed in Facebook's estimates with, say, a lot of migrants from country X in location A, but relatively few in location B, are very close to official statistics. Palotti et al. (2020) also found a correlation coefficient of 0.99 when comparing estimates for the number of Venezuelan migrants and refugees across 17 host countries as reported in a 2019 R4V[14] report with Facebook estimates for the same group of countries.

Apart from the good spatial similarity to ground truth data, the real-time value of Facebook data has also been demonstrated in the setting of an exodus after a natural disaster or economic crisis (Spyratos et al., 2019; Palotti et al., 2020; Alexander et al., 2019). Spyratos et al. (2019) and Palotti et al. (2020) looked at the case study of Venezuelan migrants and refugees after the economic crisis to study their distribution in neighbouring countries. Both groups were able to successfully detect an increase in Venezuelan migrants and refugees in surrounding countries after the crisis hit. On the other hand, Alexander et al. (2019) analysed outmigration from Puerto Rico in the aftermath of Hurricane Maria in 2017. As these studies have shown, Facebook data allow researchers to obtain more recent data, enabling us to carry out research quickly, especially during humanitarian crises. Another added value of Facebook is that collecting on-the-ground data is possible even during a period when in-person data collection is difficult to pursue (Perrotta et al., 2021; Pötzschke and Braun, 2017), not to mention that the cost and logistical effort are considerably less compared to traditional approaches.

[11] More information is available at https://developer.twitter.com/en/developer-terms/agreement-and-policy.

[12] More information is available at https://scholarslab.github.io/learn-twarc/06-twarc-command-basics#dehydrated-and-rehydrated-data-sets.

[13] More information is available at https://investor.fb.com/investor-news/press-release-details/2020/Facebook-Reports-Third-Quarter-2020-Results/default.aspx (accessed September 2020).

[14] More information is available at https://r4v.info/.

While only a limited number of studies have been conducted on the bias in Twitter data, more evidence exists on the bias in Facebook data since Facebook-derived estimates can be compared with traditional statistics directly. Since information on individuals' socioeconomic status is also available in most countries, some studies have taken into account the differences in penetration rates by age and sex across countries to fully understand the origin of the bias.

For instance, Spyratos et al. (2019) considered differences in penetration rates by age and gender both in the country of current residence and in the country of previous residence. Palotti et al. (2020) found that Facebook tends to overestimate the number of migrants and refugees compared to official statistics – which might themselves be underestimated. The same study also showed evidence that Facebook data might underestimate the number of migrants and refugees in less wealthy areas. The measurement of wealth was inferred by looking at the prices of devices used to access Facebook. Zagheni et al. (2017) observed that Facebook generally overestimates migrant stocks in younger age groups and underestimates stocks in older age groups. Last but not least, Ribeiro et al. (2020) analysed biases across several dimensions including race, income level and age – at different geolocation levels including country, state and city. They found that bias exists mainly towards young people and women but also towards college graduates. However, Facebook tends to capture trends relatively similar to survey data when looking at individuals with high school or graduate school levels of education and different income levels.

The Facebook advertising platform[15] was originally designed for purposes of targeted marketing to allow advertisers to select and reach a specific audience. However, this tool has also shown its utility in studying migration. Different from Twitter, Facebook provides explicit criteria that enable advertisers to selectively show ads to users that are likely migrants. Concretely, Facebook provides options to selectively show ads to any Facebook user who "used to live in [country X]" (expats), "lives abroad" and "recently moved". Most studies, described further below, use the targeting criterion "used to live in [country X]" as a proxy for migration history. The Facebook advertising platform then provides an estimate of the size of the audience matching the chosen criteria. These estimates include numbers for both monthly and daily active users. To protect user privacy, the returned estimates are rounded to two significant digits. Furthermore, user groups smaller than 1,000 users are indistinguishable from 0, as FB never returns estimates of monthly active users smaller than 1,000.

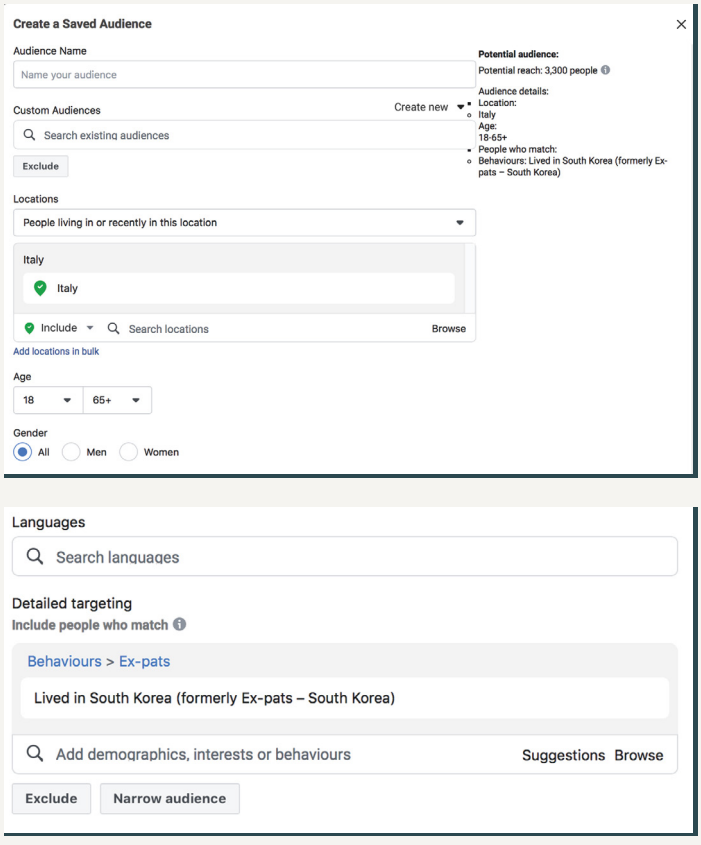
One of the key challenges for researchers is that Facebook does not disclose the methodology used to produce these estimates. For example, "expats" are briefly defined as users "who used to live in [country X] who now live abroad", but how previous countries of residence are inferred is unclear. Furthermore, this vague definition leaves us to wonder whether it closely corresponds to the definition of "migrant", as set in the United Nations 1998 Recommendations or in the revised version by the United Nations Statistical Commission. The first states that an international migrant is "any person who changes his or her country of usual residence" (DESA, 1998:9). On the other hand, the revised definition specifies that "a person who has changed his or her country of residence and established new residence in the country within a given year" (UNSD, 2020:13) is considered as an international migrant.

[15] More information is available at www.facebook.com/business/ads.

However, from the work of Herdağdelen et al. (2016), done by researchers at Facebook, we are able to get clues about the features that likely contribute to the identification of expats on the site. They first identify individuals who are now living in the United States of America but who "specify a hometown outside of the United States", their country of interest. To further reduce noise due to users not stating their true hometowns, the researchers have also considered social network information to determine if a particular hometown is plausible. At the time of writing, Facebook supports 89 countries of origin of expats. Aside from migration estimates, the Facebook ads platform provides users' demographic information such as their self-declared level of education, field of study, schools/universities attended, field or industry, and also behaviours and interests. Behaviours include various categories such as "frequent travellers" and likely engagements of individuals in conservative, liberal or moderate political contents in the United States.

A particularly useful targeting attribute is the device type and the operating system used to access Facebook. Previous works have shown that, as a first approximation, Apple iOS devices are more common among population groups with higher disposable incomes (Fatehkia et al., 2020; Palotti et al., 2020). The targetable interests include various domains, such as foods/drinks, fashion, entertainment, sports and technology – and others which are observed through pages that users "liked". Not all targeting attributes are available in all countries, with the richest set of targeting capabilities found in the United States. All of these data can be accessed using either the Facebook Ads Manager or Marketing API.[16]

Figure 2. An example of a Facebook ads platform audience search



The Facebook advertising platform was originally designed for purposes of targeted marketing to allow advertisers to select and reach a specific audience. However, this tool has also shown its utility in studying migration.

[16] More information is available at https://developers.facebook.com/docs/marketing-apis/.

As with Twitter, using Facebook as a data source also comes with several drawbacks. First of all, there are bots and fake accounts that may interfere with how Facebook identifies expats. According to Rosen (2020), there were about 1.5 billion fake accounts in the second quarter of 2020. Fortunately, most of these accounts are detected and removed almost immediately after their creation so that they account for approximately "only" 5 per cent of the worldwide monthly active users (MAU) at any given time (ibid.). Under Facebook's policy, misrepresentation of one's identity on a profile classifies as having a fake account.[17] This includes any personal profiles created to represent pets, businesses or organizations – or profiles managed by multiple users. Owning multiple accounts is also not allowed by Facebook's policy. As for spam, there were about 1.4 billion spam activities detected on Facebook during the same period. Spam activities involve inflating posts, likes or shares that may mislead audiences typically for financial gains.[18]

While with Twitter, researchers have the flexibility to implement their own algorithm to detect bots and fake accounts, this is not possible with Facebook as no anonymized individual data is shared and the exact algorithms employed to estimate the size and composition of user groups are not fully disclosed. In addition, we can extract audience estimates of Facebook only at the point of querying the API. While this makes perfect sense for advertisers, who are not interested in how many users they could have reached, say, one year ago, this limits research on historical migration trends.

Conclusion

There are both advantages and drawbacks in using social media data to study migration events. Twitter and Facebook data are freely available through public APIs, but Twitter requires considerably more API calls as individual – not aggregate – data needs to be collected. Collecting large amounts of Twitter data is also becoming more and more challenging due to increased scrutiny, resulting in regular changes in rate limits. However, once obtained, Twitter data are relatively easier to interpret. Facebook data are more difficult to interpret as the exact algorithms used to extract the estimates are not disclosed. Also, the methods periodically change, resulting in discontinuities that need to be accounted for. With Twitter data, longitudinal studies are possible as Twitter supports the collection of historical data. This is not feasible with real-time Facebook advertising data.

Although the privacy issues were not treated in depth in this chapter, they constitute one of the crucial limitations of Twitter data. Twitter is an open data platform, but users are not necessarily aware of researchers analysing their conversations and activities. To manage privacy issues, it is necessary to take security measures such as pseudonymization, or anonymization, to prevent reidentification of individuals and protect personal data. Furthermore, it is essential for researchers to secure the collected data to ensure that both raw and processed data are used responsibly and ethically. This would involve storing data in a secured server or limiting access to them. All of these privacy issues are of a lesser concern with Facebook data as the site provides only estimates of migrants at an aggregate level. At the same time, questions of potential group harm are more pronounced as it is relatively easy to identify, or rather enumerate, potentially vulnerable groups, based on variables such as income or ethnicity.

[17] More information is available at www.facebook.com/communitystandards/misrepresentation.

[18] More information is available at www.facebook.com/communitystandards/spam.

The quality of migration data varies vastly from one country to another, and for those with low-quality data, the need for better data has been pointed out. In such data-poor environments, social media data could help augment traditional ways of data collection to create better migration statistics. However, for countries with strong and real-time registration systems, social media is unlikely to add meaningful value in terms of providing better statistics. But even in these settings, innovative data might be able to complement existing data by adding new dimensions. For example, it seems promising to augment the Integration Index that is traditionally developed using civil registration data (Bansak et al., 2018). This index could potentially be improved by adding new insights from social media data, such as spatial and cultural aspects of integration (Dubois et al., 2018; Mazzoli et al., 2020; Stewart et al., 2019) – where, in addition to traditional measures of culture (such as languages, religion or marital status), "likes" or interests of individual users, as they relate to different aspects of life, can be studied through Facebook data.

REFERENCES*

Alexander, M., K. Polimis and E. Zagheni

2019 The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data. *Population and Development Review*, 45(3):617–630.

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence and J. Weinstein

2018 Improving refugee integration through data-driven algorithmic assignment. *Science*, 359(6373):325–329.

Dubois, A., E. Zagheni, K. Garimella and I. Weber

2018 Studying migrant assimilation through Facebook interests. Accepted as a short paper at the International Conference on Social Informatics. Springer, pp. 51–60.

Fatehkia, M., I. Tingzon, A. Orden, S. Sy, V. Sekara, M. Garcia-Herranz and I. Weber

2020 Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, 9:22.

Fiorio, L., G. Abel, J. Cai, E. Zagheni, I. Weber and G. Vinué

2017 Using Twitter data to estimate the relationship between short-term mobility and long-term migration. In: *WebSci '17: Proceedings of the 2017 ACM on Web Science Conference*. Association for Computing Machinery (ACM), New York, pp. 103–110.

Gendronneau, C., A. Wiśniowski, D. Yildiz, E. Zagheni, L. Fiorio, Y. Hsiao, M. Stepanek, I. Weber, G. Abel and S. Hoorens

2019 *Measuring Labour Mobility and Migration Using Big Data: Exploring the Potential of Social-Media Data for Measuring EU Mobility Flows and Stocks of EU Movers*. European Commission, Brussels. Available at www.rand.org/pubs/external_publications/EP68038.html.

Hausmann, R., J. Hinz and M.A. Yildirim

2018 Measuring Venezuelan emigration with Twitter. Kiel Working Paper No. 2106. Kiel Institute for the World Economy.

Hawelka, B., I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos and C. Ratti

2014 Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

Herdağdelen, A., B. State, L. Adamic and W. Mason

2016 The social ties of immigrant communities in the United States. In: *Proceedings of the Eighth ACM Conference on Web Science*. ACM, New York, pp. 78–84.

** All hyperlinks were working at the time of writing this report.*

Hsiao, Y., L. Fiorio, J. Wakefield and E. Zagheni
2020 Modeling the bias of digital data: An approach to combining digital and survey data to estimate and predict migration trends. MPIDR Working Paper WP-2020-019. Max Planck Institute for Demographic Research, Rostock, Germany.

Huang, X., Z. Li, Y. Jiang, X. Li and D. Porter
2020 Twitter reveals human mobility dynamics during the COVID-19 pandemic. *PLOS ONE*, 15(11):e0241957.

Huang, W., I. Weber and S. Vieweg
2014 Inferring nationalities of Twitter users and studying inter-national linking. In: *HT '14: Proceedings of the 25th ACM Conference on Hypertext and Social Media*. ACM, New York, pp. 237–242.

Kim, J., A. Sîrbu, F. Giannotti and L. Gabrielli
2020 Digital footprints of international migration on Twitter. In: *International Symposium on Intelligent Data Analysis*. Springer, pp. 274–286.

Lamanna, F., M. Lenormand, M.H. Salas-Olmedo, G. Romanillos, B. Gonçalves and J.J. Ramasco
2018 Immigrant community integration in world cities. *PLOS ONE*, 13(3):e0191612.

Mazzoli, M., B. Diechtiareff, A. Tugores, W. Wives, N. Adler, P. Colet and J.J. Ramasco
2020 Migrant mobility flows characterized with digital data. *PLOS ONE*, 15(3):e0230264.

Morstatter, F. and H. Liu
2017 Discovering, assessing, and mitigating data bias in social media. *Online Social Networks and Media,* 1:1–13.

O'Brien, M.L. and M.A. Eger
2020 Suppression, spikes, and stigma: How COVID-19 will shape international migration and hostilities toward it. *International Migration Review*, 55(3):640–659.

Palotti, J., N. Adler, A. Morales-Guzman, J. Villaveces, V. Sekara, M. Garcia Herranz, M. Al-Asad and I. Weber
2020 Monitoring of the Venezuelan exodus through Facebook's advertising platform. *PLOS ONE*, 15(3):e0230455.

Perrotta, D., A. Grow, F. Rampazzo, J. Cimentada, E. Del Fava, S. Gil-Clavel and E. Zagheni
2021 Behaviors and attitudes in response to the COVID-19 pandemic: Insights from a cross-national Facebook survey. *EPJ Data Science*, 10:17.

Pötzschke, S. and M. Braun
2017 Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. *Social Science Computer Review*, 35(5):633–653.

Ribeiro, F.N., F. Benevenuto and E. Zagheni

2020 How biased is the population of Facebook users? Comparing the demographics of Facebook users with census data to generate correction factors. In: *WebSci '20: 12th ACM Conference on Web Science*. ACM, New York, pp. 325–334.

Rosen, G.

2020 Community standards enforcement report: August 2020. Facebook. Available at https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/.

Sîrbu, A., G. Andrienko, N. Andrienko, C. Boldrini, M. Conti, F. Giannotti, R. Guidotti, S. Bertoli, J. Kim, C.I. Muntean, L. Pappalardo, A. Passarella, D. Pedreschi, L. Pollacci, F. Pratesi and R. Sharma

2020 Human migration: The big data perspective. *International Journal of Data Science and Analytics*, 11:341–360.

Spyratos, S., M. Vespe, F. Natale, I. Weber, E. Zagheni and M. Rango

2018 *Migration Data Using Social Media: A European Perspective*. Publications Office of the European Union.

2019 Quantifying international human mobility patterns using Facebook network data. *PLOS ONE*, 14(10):e0224134.

Stewart, I., R.D. Flores, T. Riffe, I. Weber and E. Zagheni

2019 Rock, rap, or reggaeton?: Assessing Mexican immigrants' cultural assimilation using Facebook data. Proceedings of the World Wide Web Conference, pp. 3258–3264.

United Nations Department of Economic and Social Affairs (DESA)

1998 *Recommendations on Statistics of International Migration*. Revision 1. New York. Available at https://unstats.un.org/unsd/publication/seriesm/seriesm_58rev1e.pdf.

2019a *International Migration 2019: Report*. ST/ESA/SER.A/438. Population Division, New York. Available at www.un.org/en/development/desa/population/migration/publications/migrationreport/docs/InternationalMigration2019_Report.pdf.

2019b *International Migration 2019: Wall Chart*. ST/ESA/SER/A/431. Population Division, New York. Available at www.un.org/en/development/desa/population/migration/publications/wallchart/docs/MigrationStock2019_Wallchart.pdf.

United Nations Statistical Commission (UNSD)

2020 Report of the Secretary-General on migration statistics. E/CN.3/2021/11. Available at https://unstats.un.org/unsd/statcom/52nd-session/documents/2021-11-MigrationStats-EE.pdf.

Zagheni, E., V.R.K. Garimella, I. Weber and B. State

2014 Inferring international and internal migration patterns from Twitter data. Proceedings of the 23rd International Conference on World Wide Web, pp. 439–444.

Zagheni, E., I. Weber and K. Gummadi

2017 Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43(4):721–734.

4

SOCIAL MEDIA

ARTIFICIAL INTELLIGENCE/ MACHINE LEARNING

OTHER

FORECASTING HUMAN MOBILITY

TRACKING HUMAN DISPLACEMENT

AN INNOVATIVE FRAMEWORK FOR ANALYSING ASYLUM-RELATED MIGRATION

Constantinos Melachrinos,[1] *Marcello Carammia*[2] *and Teddy Wilkin*[3]

Introduction

Asylum-related migration comprises those migrants who left their countries of origin with the intention to seek international protection in another country or ended up doing so. People with legitimate needs for international protection, according to the 1951 Refugee Convention and its 1967 Protocol, and the Qualification Directive (2011/95/EU) (European Union, 2011), as well as people fleeing their countries for better economic/ employment opportunities are included, as long as they (intend to) seek asylum. Moreover, everyone has the right to apply for asylum, irrespective of whether they arrived in any given country regularly or irregularly.

Asylum-related migration is therefore a complex process and one difficult to typify. Yet understanding migration processes and their drivers is key to evidence-based policymaking, while some capacity to anticipate migration flows (to the extent possible, see: Carammia and Dumont, 2018) is crucial to the design of efficient[4] asylum systems. This chapter presents the European Union Agency for Asylum (EUAA) approach to the analysis of asylum-related migration. It first provides a description of the asylum data exchanged under the EUAA Early warning and Preparedness System (EPS). Subsequently, it describes the EUAA analytical and conceptual framework, discussing the analytical purpose, temporal dimension, resources required, and the stakeholders that define the target audience for the EUAA analytical products.

[1] Constantinos Melachrinos is Senior Analyst at the European Union Agency for Asylum (EUAA). He holds an undergraduate degree in Physics from the University of Chicago and a PhD in Physics from the Massachusetts Institute of Technology.

[2] Marcello Carammia is Senior Researcher at the Department of Political and Social Sciences of the University of Catania, Italy. Marcelo is a founding co-director of the Italian Agendas Project and a principal investigator of the European Union Policy Agendas Project. Previously he worked as Senior Researcher at the European Asylum Support Office (EASO) (now EUAA).

[3] Teddy Wilkin is Head of the Data Analysis and Research Sector at EUAA. Previously he was a Research Fellow at the University of Oxford. He also holds a PhD from Oxford.

[4] In this context, efficiency is not about cost savings, but about foreseeing sudden increases in caseload to allow countries to have resources in place to process each application quickly and in line with international standards.

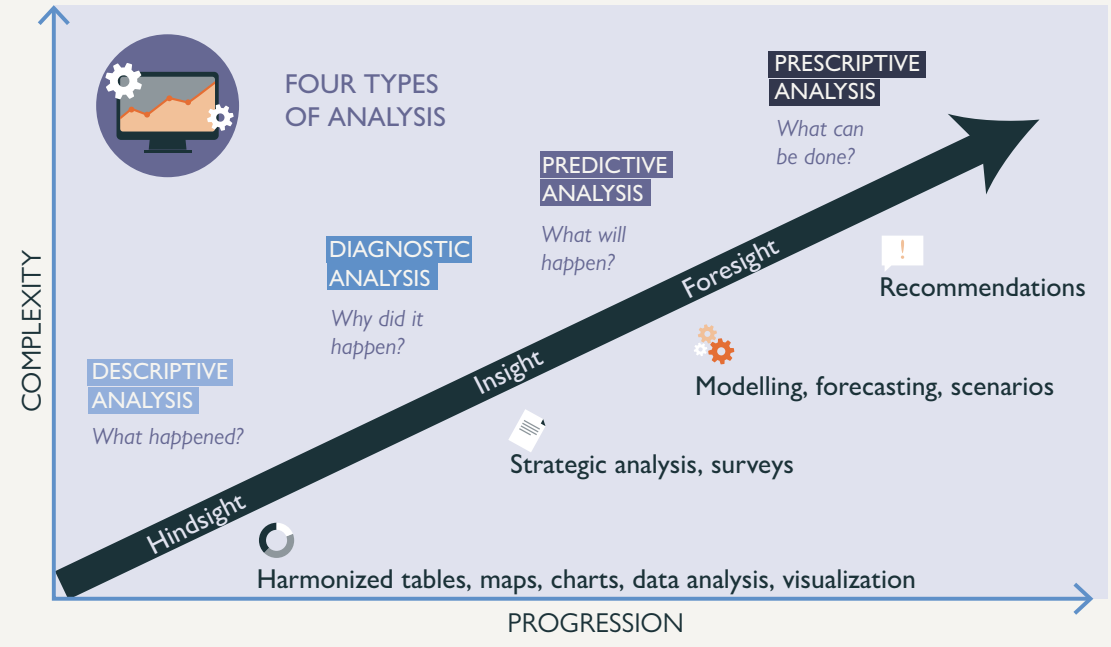
Early warning and Preparedness System (EPS) data

In 2012, to better understand asylum-related migration to the European Union, Norway and Switzerland (hereafter referred to as EU+ countries), EUAA[5] launched its EPS, an information-exchange mechanism aimed at providing EUAA, European Union member States, Schengen associated countries and the European Commission with timely, accurate, and comparable data on the lodging and processing of asylum applications in EU+ countries, in line with the Common European Asylum System.[6]

Quantification of complex real-world phenomena invariably involves a degree of simplification. Hence, EPS data are extensive and have been shown to add much value, but they also have their limitations. Consistent with most data originating from administrative bodies, EPS data tend to count and describe administrative procedures rather than actual people, plus at the supranational level, it is not yet possible to follow individual asylum cases as they progress through (and often between) convoluted asylum procedures,[7] nor is it possible at the EU+ level to link individual asylum applications to other administrative procedures such as regular (e.g. visa applications, resettlement) or irregular migration (e.g. refusals of entry, detections of illegal border-crossing or illegal stay).

That said, EPS data are the cornerstone of the EUAA analytical framework, and much intellectual enterprise has been spent designing the data and making sure that the indicators add as much value as possible to the four types of analytics – descriptive, diagnostic, predictive and prescriptive – as well as addressing different time horizons, including nowcasting and monitoring, early warning, and forecasting.

Figure 1. Four types of analysis



[5] Before January 2022, this was known as EASO.

[6] See: EASO, 2011.

[7] Interoperability Regulations (EU) 2019/817 and 2019/818, and proposals of the European Union Pact on Migration and Asylum 2020, take significant steps to mitigate this data limitation.

Nowcasting

The assessment of the asylum situation relies largely on backward-facing analyses using EPS indicators, data shared by partner organizations, and publicly available information. EPS indicators include data on all stages of the asylum procedure,[8] including access to procedure, reception systems, first instance determination, returns and the Dublin Regulation. This assessment provides for a retrospective understanding of events and uses the four types of analytics to speculate about their drivers. Embedded in this logic is the assumption that an effective description of the recent past can also say something about the present and our expectations for the near future (Banbura et al., 2010). Repeated and regular analyses are a crucial part of any analytical framework because reoccurring trends and stable drivers can underpin predictive analytics, or at least can be used to corroborate their outputs. At EUAA, weekly, monthly and biannual analyses of the current asylum situation are systematically produced, and the findings communicated to internal and external stakeholders, including European Union and national policymakers, to guide their response to recent and ongoing events. These analyses take the form of descriptive and diagnostic analytics, as mentioned in the previous section, where past events are described and interpreted through the prism of strategic analysis. For example, monitoring the situation in the same or neighbouring destination countries over time, or comparing different countries during the same period of time, or a combination of the two approaches, can elucidate why the asylum situation in EU+ countries is what it is, leading to valuable inputs for predictive and prescriptive analytics.

Analyses connect current and past trends. While past events may provide a guide and a framework for future expectations, it is usually the case that drivers of asylum-related migration vary between individual migration flows or events, and can even be so short-lived that drivers vary within a single migration flow. For example, increases in conflict events in the Syrian Arab Republic may lead to an increase of asylum-related migration into neighbouring Türkiye, but not to EU+ countries, whereas increases in conflict events in Libya may lead to increases in asylum-related migration directly into EU+ countries (i.e. into Italy or Malta). On the other hand, before the European Union–Türkiye statement, conflict events in the Syrian Arab Republic led directly to higher numbers of incoming asylum applicants in EU+ countries, which is not the case anymore. As such, the results of any analyses should be interpreted along with the caveat that the present and future situations may not always be explainable using the recent past, especially in cases of extreme shocks to the system. For example, the ongoing COVID-19 pandemic and the subsequent travel restrictions and generally reduced mobility have resulted in fewer asylum applications being lodged in EU+ countries despite the fact that more people globally are in need of international protection, and it has disrupted many long-standing migration patterns, such as the Global South to North and the rural–urban spectrums.

[8] More information on the EUAA's Early warning and Preparedness System indicators is available at www.euaa.europa.eu/asylum-knowledge/data-analysis-and-research.

Monitoring

As a first step towards early warning of asylum-related migration, EUAA has developed a system that continuously monitors the administrative data (EPS indicators), as well as novel sources of data, such as Internet searches for asylum-related migration topics (Google, n.d.) and potential "push factor" events in countries of origin of asylum seekers, extracted from event-based databases using global media reports, such as the GDELT Project (n.d.) and the Armed Conflict Location and Event Dataset (Raleigh et al., 2010).

Administrative data, such as the number of asylum applications lodged in each EU+ country, broken down by citizenship, may exhibit trends that allow for effective nowcasting of the overall asylum situation. The proportion of asylum decisions that grant (as opposed to refuse) international protection (European Union-regulated recognition rate) in turn is positively associated with asylum-related migration (Toshkov, 2014; Brekke et al., 2017), assuming that migrants prefer to go to destinations where they are more likely to be granted international protection and that this information is available to them; it can also deflect potential asylum seekers into irregularity, based on an analysis of bilateral asylum and visa policies on migrant flows to 29 European States in the 2000s (Czaika and Hobolth, 2016).

Internet searches for asylum-related topics, in addition to searches for specific countries of transit or destination, are downloaded and monitored by EUAA because they have been shown to indicate the intentions of potential asylum seekers to migrate (Böhme et al., 2020). Monitoring Internet searches and attributing intentions of migration should be performed with caution since the link may not always exist, or intentions to migrate may not possibly materialize due to extraneous factors. Algorithms behind Internet searches may also affect the results, necessitating a thorough understanding of the underlying methodologies, before venturing into this kind of analysis.

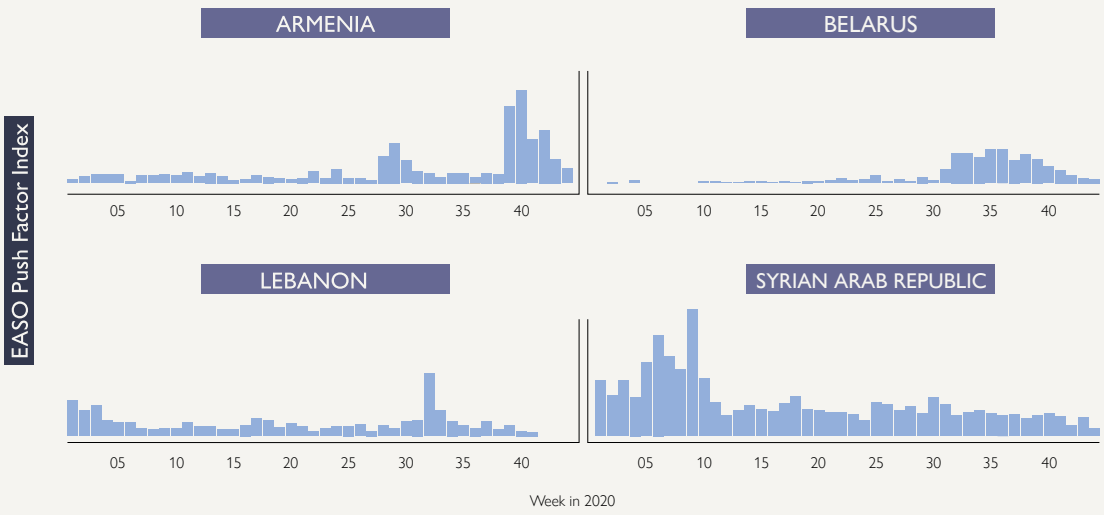
The GDELT Project extracts from global media near to real-time data on events around the world, classified automatically into about 200 categories (Gerner et al., 2002). EUAA has identified which events have the potential to generate asylum-related migration and aggregated them into five macrocategories – conflict, economic, political, governance and social – reflecting broad drivers or push factors in the countries of origin/transit of potential asylum seekers. The selected events are also weighed according to their severity and relevance to migration to construct the Push Factor Index (PFI). The PFI is updated daily and visualized in interactive dashboards, representing in a snapshot the situation in the countries of origin, a tool that is useful for migration practitioners in the field, as well as for policymakers. One advantage of the GDELT Project is the ability to monitor the situation in countries of origin and transit at the subnational level, since each event is automatically attributed to a specific geographical location. The fact that geolocation and attribution to Conflict and Mediation Event Observations (CAMEO) categories is performed automatically by the GDELT Project algorithms raises the need to critically evaluate these elements and understand potential limitations or biases (Hammond and Weidmann, 2014). For example, when a location is not provided in an article, an event is often attributed to the capital of the country where the event took place, even if the event took place somewhere else, creating a "capital city" bias in these results.

Clearly, the relationship between events in countries of origin and asylum-related migration is not straightforward: applying for asylum in EU+ countries is only one of the possible outcomes of crises in third countries. However, monitoring the PFI does provide important insights into migration processes, and it also bears the potential for early warning purposes. A recent study (Melachrinos et al., 2020) has shown that the PFI in countries in Africa was correlated with the number of asylum applications by nationals to EU+ countries during the years 2016 and 2017. In subsequent years (2018–2019), the relationship is less strong, possibly because of fewer search and rescue operations in the Central Mediterranean route and increased interceptions by the Libyan Coast Guard and Navy of migrants attempting to cross the route from Libya to Italy, one of the most important routes into Europe for African nationals seeking international protection in the European Union.

Figure 2 shows the weekly PFI in four countries that were of interest for asylum-related migration in 2020: Armenia, Belarus, Lebanon and the Syrian Arab Republic. The recent conflict in Artsakh/Nagorno–Karabakh, which restarted on 27 September 2020 (*BBC News*, 2020a), is reflected in a corresponding peak of PFI. Push factors have decreased since the agreed ceasefire, but at the time of writing, it was not known whether this event would lead to increased numbers of asylum applications in EU+ countries by Armenian citizens or whether they will be internally displaced or seek asylum in neighbouring countries. Other recent examples where migration drivers have intensified in countries of origin, and have resulted in an increase in PFI, include the blast in Beirut in August 2020 (Hubbard and El-Naggar, 2020), the contested elections in Belarus (*Al Jazeera*, 2020) and subsequent demonstrations in August/September 2020. Equivalently, reductions in migration drivers, such as those caused by the recent agreement for a ceasefire in the Syrian Arab Republic in March 2020 (*BBC News*, 2020b), are reflected in a decrease in PFI and stabilization at lower levels, which may lead to lower numbers of asylum seekers in the European Union.

Figure 2. The Push Factor Index in Armenia, Belarus, Lebanon and the Syrian Arab Republic weekly in 2020

The figure shows the weekly PFI in four countries that were of interest for asylum-related migration in 2020: Armenia, Belarus, Lebanon and the Syrian Arab Republic.



Source: EUAA and the GDELT Project.

Early warning alerts

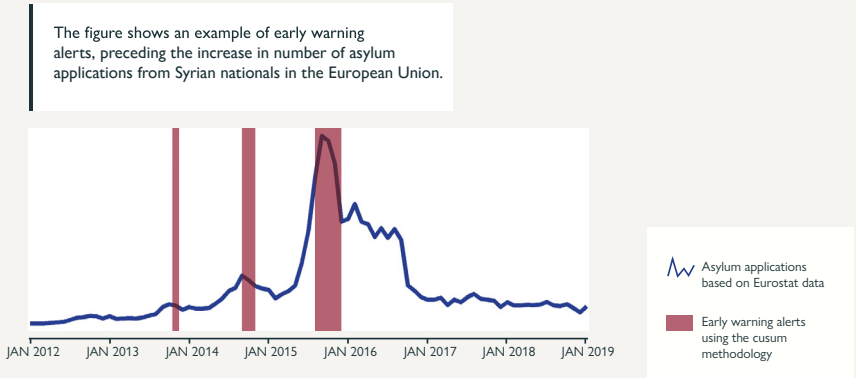
While monitoring is useful to discover trends and observe events unravelling more or less as they occur, it still depends on trained analysts to visualize the data, manually assess the asylum-related situation, and place trends in the correct context. Alternatively, the cumulative sum approach is an automated method that generates early warning alerts based on statistical control theory (Bijak et al., 2017).

The cumulative sum (or "cusum") method, introduced by E.S. Page (1954), monitors the cumulative sum of some function of the observed data – in this case, the number of asylum applications. Hence the cumulative sum can generate an early warning alert when it exceeds a certain threshold which can be set based on the past performance of the indicator or by using a formal model.

The acceptance thresholds to trigger alerts depend on a moving average window of the latest data available for each country of origin. Single countries of origin have different "background" levels of conflicts and other potential migration-generating events, have different patterns of Internet searches, generate different volumes of asylum applicants, and so forth. For this reason, fixed thresholds may result in inconsistent "false positive" (Type I error) alarms or "false negative" (Type II error) outcomes. The trade-off between false positives and false negatives ultimately depends on policy considerations and the tolerance towards each of the two errors.

Figure 3 shows a time series of the number of asylum applications lodged by Syrian nationals in EU+ countries between 2012 and 2018, based on Eurostat data. The early warning alerts, shown in shaded areas, were triggered using the cusum methodology, often preceding a subsequent increase in the number of asylum applications. Different methods for alerts, such as the statistical significance of the distance from a moving average, would generate different alerts, and so all such methods are to some extent subjective. The sensitivity of the alerts and the extent to which there are false positives depend on the activation threshold for triggering alerts. This is a preliminary application of the method, adapted from Napierała et al. (2021), with improvements planned to enable earlier and more relevant alerts.

Figure 3. Example of application of cusum to data on the monthly number of asylum applications lodged in the European Union by Syrian nationals (Eurostat data), with early warning alerts triggered (shaded areas)



Cusum is one possible way of monitoring time series, which EUAA combines with other approaches such as the analysis of short-term and medium-term moving averages in the series borrowed from financial econometrics (Murphy, 1999).

While individual early warning alerts are useful to capture the attention of analysts and point them to specific time series that exhibit sharp changes from their expected values, the concurrent change in a number of early warning alerts for specific countries can highlight that more than one time series are affected by recent events. For example, an economic embargo in a country of origin can increase the economic drivers, increasing the PFI sharply and generating an early warning alert, which is complemented by an early warning alert elicited by Internet searches for asylum and visa in a specific European Union country or country of transit. This example illustrates that not only the migration drivers but also the intentions to migrate have increased, and the system is able to automatically generate such early warning alerts for further investigation by analysts and subsequent action by policymakers.

Forecasting

Asylum-related migration is a complex process driven by a multitude of factors which themselves depend on external factors – such as climate, political events and crises, and State actions such as policy (asylum law, migration management) – and ultimately on individual actions such as the decision to migrate (Bijak, 2011; Bijak et al., 2017). EUAA has recently developed an early warning and forecasting system (EPS–Forecasting) which uses an adaptive elastic-net model that combines data on events and Internet searches in third countries, irregular crossings at the European Union border, and recognition rates in EU+ countries. By training the model in the recent past, short-term forecasts of bilateral asylum-related migration flows, indicated by the number of asylum applications by nationals from specific countries of origin to European Union member States, are generated. Further information about the EUAA early warning and forecasting system can be found in the article by Carammia et al. (2022).

Requirements

Stakeholders using the EUAA analyses to guide their response to the asylum situation have different requirements that stem from their work mandates. These stakeholders range from operations, asylum and reception services in member States, to policymakers at the European Union and national levels. Their varying requirements are outlined below:

(a) Operations need short-term information to support decision-making at the tactical level (i.e. where to deploy human and material resources to address challenges imposed by sudden changes in asylum-related migration). These include border guards, registration officers and vulnerability experts.
(b) Asylum and reception services have planning requirements for the medium term. They seek this information in order to be able to employ the most appropriate number of caseworkers to process asylum applications, to ensure that they have the right capacity to accommodate asylum applicants and more.

(c) Policymakers need to respond in longer time periods, following stable European Union or national-level trends. Data used by policymakers should be integrated so that strong policy developments in one area do not cause displacement effects to a weaker policy in another area. For example, the adoption of a list of "safe" countries of origin in one European Union member State may drive applicants from those safe countries to seek asylum in another member State, leading to an increase in secondary movements within the European Union.

The different requirements among the stakeholders give rise to the need for different outputs, which consider the analyses performed to address the asylum situation from the perspective of these stakeholders.

Outputs

To address the challenges faced by operational colleagues, short-term forecasts and early warning alerts are needed. These outputs, included in early warning reports disseminated to operational stakeholders in a timely manner, allow for an understanding of the current situation and how it is expected to evolve in the following few weeks, with the aim of improving operational planning. Asylum and reception services employ qualitative reports, which analyse whether current trends are likely to continue in the medium term. These reports improve planning for human resources and accommodation for asylum applicants. Policymakers, on the other hand, need to understand the long-term trends in migration drivers and the relationships between those drivers. Investigating the causal relationships between migration drivers and administrative indices allows policymakers to adjust policies in a way that would improve the asylum processes to ensure safe and efficient access to international protection for those in need, while at the same time also alleviating the strain of sudden large increases in the number of asylum applicants on national reception systems.

Resources

Considering these requirements by stakeholders, it is imperative that results are communicated to them in a timely manner. However, the dissemination of raw quantitative results is discouraged due to the high risk of misinterpretation and overconfidence in quantitative approaches. Data science, the science of uncovering insights from data, can generate meaningful results, which, however, need to be validated by domain-expert analysts. Only after this validation and interpretation of the results by analysts can the main findings be simplified and shared with stakeholders. This validation process also ensures that any caveats or issues introduced due to missing data or inconsistencies arising from data reporting by different countries can be communicated clearly alongside the main findings. Since the results of such analyses can have an impact on the availability of national or international forms of protection, it is important to include this additional layer of interpretation before the results can be used to affect operational resource planning or policymaking (Albertinelli et al., 2020).

Conclusions

Policymakers and operational, asylum, and reception services employ the analysis of asylum-related migration developed by EUAA to understand the main drivers, as well as design or adjust their response to ever-changing trends. As a result, the output portfolio of EUAA has been designed to address a variety of challenges, with different time horizons and utilizing different methodologies depending on the requirements of each stakeholder. The need for domain-expert analysts to validate and filter the raw results of the data science analysis presented earlier is important, especially since the reports may be used to alter policies affecting human lives.

REFERENCES*

Albertinelli, A., P. Alexandrova, C. Melachrinos and T. Wilkin

2020 Forecasting asylum-related migration to the European Union, and bridging the gap between evidence and policy. *Migration Policy Practice*, X(4):35–41. Available at https://publications.iom.int/books/migration-policy-practice-vol-x-number-4-september-december-2020.

Al Jazeera

2020 Tens of thousands hold biggest protest yet, Lukashenko defiant. 17 August. Available at www.aljazeera.com/news/2020/8/17/tens-of-thousands-hold-biggest-protest-yet-lukashenko-defiant.

Banbura, M., D. Giannone and L. Reichlin

2010 Nowcasting. *ECB Working Paper No. 1275*. Available at https://ssrn.com/abstract=1717887.

BBC News

2020a Armenia and Azerbaijan fight over disputed Nagorno–Karabakh. 28 September. Available at www.bbc.com/news/world-europe-54314341.

2020b Syria war: Russia and Turkey agree Idlib ceasefire. 5 March. Available at www.bbc.com/news/world-middle-east-51747592.

Bijak, J.

2011 *Forecasting International Migration in Europe: A Bayesian View*. Springer, Dordrecht. Available at https://link.springer.com/book/10.1007/978-90-481-8897-0.

Bijak, J., J.J. Forster and J. Hilton

2017 *Quantitative Assessment of Asylum-Related Migration: A Survey of Methodology*. European Asylum Support Office, Malta. Available at https://data.europa.eu/doi/10.2847/642161.

Böhme, M.H., A. Gröger and T. Stöhr

2020 Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142.

Brekke, J.P., M. Røed and P. Schøne

2017 Reduction or deflection? The effect of asylum policy on interconnected asylum flows. *Migration Studies*, 5(1):65–96.

** All hyperlinks were working at the time of writing this report.*

Carammia, M. and J.C. Dumont

2018 Can we anticipate future migration flows? OECD Migration Policy Debates No. 16, May. Available at www.oecd.org/els/mig/migration-policy-debate-16.pdf.

Carammia, M., S.M. Iacus and T. Wilkin

2022 Forecasting asylum-related migration flows with machine learning and data at scale. *Scientific Reports*, 12:1457.

Czaika, M. and M. Hobolth

2016 Do restrictive asylum and visa policies increase irregular migration into Europe? *European Union Politics*, 17(3):345–365.

European Asylum Support Office (EASO)

2011 Work Programme 2012. EASO/MB/2011/25. Available at https://euaa.europa.eu/sites/default/files/easo_2011_00110000_en.pdf.

European Union

2011 Directive 2011/95/EU of the European Parliament and of the Council of 13 December 2011 on standards for the qualification of third-country nationals or stateless persons as beneficiaries of international protection, for a uniform status for refugees or for persons eligible for subsidiary protection, and for the content of the protection granted (recast). *Official Journal of the European Union*. Available at https://eur-lex.europa.eu/eli/dir/2011/95/oj.

The GDELT Project

n.d. Intro – Watching our world unfold: A global database of society. Available at www.gdeltproject.org/ (accessed November 2020).

Gerner, D.J., A.J. Rajaa, P.A. Schrodt and Ö. Yilmaz

2002 Conflict and Mediation Event Observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. Paper presented at the International Studies Association, New Orleans, March.

Google

n.d. Google Trends. Available at https://trends.google.com/ (accessed November 2020).

Hammond, J. and N.B. Weidmann

2014 Using machine-coded event data for the micro-level study of political violence. *Research and Politics*, 1(2).

Hubbard, B. and M. El-Naggar

2020 Clashes erupt in Beirut at blast protest as Lebanon's anger boils over. *The New York Times*, 8 August. Available at www.nytimes.com/2020/08/08/world/middleeast/Beirut-explosion-protests-lebanon.html.

Melachrinos, C., M. Carammia and T. Wilkin

2020 Using big data to estimate migration "push factors" from Africa. In: *Migration in West and North Africa and across the Mediterranean: Trends, Risks, Development and Governance* (Fargues, F., M. Rango, E. Börgnas and I. Schöfberger, eds.). IOM, Geneva, pp. 98–116. Available at https://publications.iom.int/books/migration-west-and-north-africa-and-across-mediterranean-chapter-8.

Murphy, J.J.

1999 *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. New York Institute of Finance, Paramus. Available at https://cdn.preterhuman.net/texts/unsorted2/Stock%20books%20029/John%20J%20Murphy%20-%20Technical%20Analysis%20Of%20The%20Financial%20Markets.pdf.

Napierała J., J. Hilton, J.J. Forster, M. Carammia and J. Bijak

2021 Toward an early warning system for monitoring asylum-related migration flows in Europe. *International Migration Review*, 56(1):33–62.

Page, E.S.

1954 Continuous inspection schemes. *Biometrika*, 41(1–2):100–115.

Raleigh, C., A. Linke, H. Hegre and J. Karlsen

2010 Introducing ACLED: An Armed Conflict Location and Event Dataset. *Journal of Peace Research*, 47(5):651–660.

Toshkov, D.D.

2014 The dynamic relationship between asylum applications and recognition rates in Europe (1987–2010). *European Union Politics*, 15(2):192–214.

ADDITIONAL READING

Acostamadiedo, E., R. Sohst, J. Tjaden, G. Groenewold and H. de Valk

2020 *Assessing Immigration Scenarios for the European Union in 2030: Relevant, Realistic and Reliable?* IOM, Geneva, and the Netherlands Interdisciplinary Demographic Institute, The Hague. Available at https://publications.iom.int/books/assessing-immigration-scenarios-european-union-2030.

Harrison, P.J. and P.P. Veerapen

1994 A Bayesian decision approach to model monitoring and cusums. *Journal of Forecasting*, 13(1):29–36.

5

ARTIFICIAL INTELLIGENCE/ MACHINE LEARNING

OTHER

FORECASTING HUMAN MOBILITY

TRACKING HUMAN DISPLACEMENT

BUILDING DATA PARTNERSHIPS

ARTIFICIAL INTELLIGENCE-BASED PREDICTIVE ANALYTICS IN THE HUMANITARIAN SECTOR: THE CASE OF PROJECT JETSON

Catherine Schneider,[1] *Rebeca Moreno Jimenez*[2] *and Sofia Kyriazi*[3]

Introduction

Understanding the factors that drive and influence forced displacement is an extremely complex process, especially when that understanding is fundamental to anticipate action on humanitarian preparedness and response. For this purpose, the Office of the United Nations High Commissioner for Refugees (UNHCR) launched an experimental project aimed at anticipating population movement by applying data science techniques, namely predictive analytics, to support the UNHCR Somalia team to leverage their wealth of data on protection monitoring and population flow. The result of this was Project Jetson (http://jetson.unhcr.org), designed to provide field operations with an estimate of the number of arrivals of refugees and internally displaced persons (IDPs) in the different Somalia regions and in the cross-border area of Dollo Ado, the southern region of Ethiopia. The combination of protracted conflict and climate anomalies affecting the region (e.g. drought or floods) had created a complex and dire situation, causing UNHCR teams to think beyond the traditional population estimation methods and explore other avenues that could predict the population flows in the different regions. The resulting tool operated at the nexus of protracted conflict, climate change and forced displacement, and served as an ongoing exploratory project of the ways in which predictive analytics could support humanitarian response.

Project Jetson was not the first predictive analytics project in UNHCR, but the first one leveraging artificial intelligence (AI) to do so. At the outbreak of the 2015 Mediterranean crisis in Europe, UNHCR created the Winter Cell to anticipate refugee movement by analysing other climate anomalies, such as wave tides and ocean currents, to anticipate

[1] Catherine Schneider is Associate Innovation Officer at the Office of the United Nations High Commissioner for Refugees (UNHCR). She holds a bachelor's degree in Global Development Studies from the University of Virginia and an MSc in International Development and Humanitarian Emergencies from the London School of Economics and Political Science.

[2] Rebeca Moreno Jimenez is Innovation Officer and Data Scientist of the UNHCR's Innovation Service. Rebeca holds a bachelor's degree in International Relations and an MPP from the Instituto Tecnológico y de Estudios Superiores de Monterrey, and an MPA focusing on the management of technology policy and humanitarian affairs from Columbia University.

[3] Sofia Kyriazi is an Artificial Intelligence Engineer at the UNHCR's Innovation Service. She holds an undergraduate degree in Computer Science and Telecommunications from the University of Athens, and a master's degree with a focus on Human Media Interaction from the University of Twente, the Netherlands.

the amount of people crossing in the Mediterranean Sea. The initiative focused on identifying the at-risk points along the routes to Europe taken by refugees as well as the points that could be negatively impacted by the coming winter, all using real-time data about weather conditions (UNHCR, 2016). The UNHCR Somalia team learned about Winter Cell efforts and wanted to expand on the idea, in a totally different operational context and with different constraints. After several rainy seasons failed between 2016 and 2017, the UNHCR Somalia team, particularly its information management (IM) team, was concerned about drought conditions in the country and the potential effect on already ongoing forced displacement. The team reached out to UNHCR Innovation Service in early 2017 to support their idea of attempting to estimate the number of potential future arrivals in the most affected regions.

UNHCR Innovation had a hypothesis to explore: to see if the same patterns of climate and conflict would have the same effect on forced displacement as observed during the 2011 drought and subsequent famine, which led to hundreds of thousands fleeing to Kenya and Ethiopia (Seal and Bailey, 2013). The Somalia team and Innovation Service agreed to try to answer a specific research question out of the many that the team had: "Could we predict in advance the number of refugee arrivals with a certain degree of accuracy?" Other questions included: *What are the preferred regions for arrival? How much time do they take to flee from one region to another?* Given the resource constraint, the teams decided to focus on one: number of arrivals.

Using a combination of traditional statistics with machine-learning techniques (e.g. artificial intelligence-based linear regression time series analysis (TSA) open source algorithms), the teams analyse historical data sets from protection monitoring exercises about the most influential factors of displacement in the region. This would become Project Jetson and set a precedent for humanitarian organizations to utilize data science and applied artificial intelligence, with a human-centred design approach, in order to proactively make decisions in displacement contexts. Since no crisis is exactly the same, Project Jetson provides important insights into applying predictive analytics in Somalia's forced displacement context, as well as broader considerations in the appropriate ethics and human rights-based approach to deployment of technologies to anticipate, understand and respond to future crises.

Project Jetson data and methods for machine-learning algorithms and understanding the decision to flee

As is the case for any artificial intelligence-based project, data are among the most important components of Project Jetson. The data sets utilized combined historical humanitarian data sets (e.g. forcibly displaced population data (UNHCR, 2014) and non-traditional data sets that are more common to the development sector (e.g. commodity prices, climate/weather anomalies data) and the peacebuilding sector (e.g. violent conflict data). Jetson was the first predictive analytics project to attempt to quantify the triple nexus; this means the humanitarian–development–peace nexus, a policy concept that envisions stronger collaboration and coordination among actors from the fields of development cooperation, humanitarian action and peacebuilding (Hövelmann, 2020).

Creating data collaboratives: Partnerships for creating value with data innovation

Project Jetson followed a data collaboratives approach. "Data collaboratives" is a term coined by New York University Tandon School of Engineering for public–private sector actors to exchange data to help solve pressing public problems (NYU Tandon, 2017). In the case of Project Jetson, data collaboratives were divided into partners in data provision and partners in modelling/technical work.

Data collaboratives in data provision: Data used in Project Jetson came from several providers representing the triple nexus work. The UNHCR Somalia field team provided population data, derived from the Protection and Return Monitoring Network, a UNHCR-led project implemented in partnership with the Norwegian Refugee Council (PRMN, n.d.). This data set gives a unique perspective of the movement from point A (point of origin) to point B (current location of data collection), the potential movement from point B (current location) to point C (future point of displacement), as well as the conditions at departure (PRMN, 2017). Other data sets were acquired via development partners (the Food and Agriculture Organization of the United Nations, Somalia Water and Land Information Management, and Food Security and Nutrition Analysis Unit teams) and open data from non-profit organizations in the peacebuilding sector (e.g. the Armed Conflict Location and Event Data Project (ACLED)). Later, the UNHCR Sub-office in Melkadida contributed to expanding Project Jetson to predict arrivals in Ethiopia by providing aggregated and anonymized data on historical refugee arrivals. Project Jetson considered that both the data providers and the technical teams working on mathematical models are considered partners in the development of the project. Data providers gathered their resources (e.g. staff time and data sets) to explain and share their data sets for achieving UNHCR Somalia's team goal: predict displacement.

Data collaboratives in modelling/technical work: Majority of United Nations agencies working on data innovation, data science or other non-traditional data approaches to solve the most pressing operational challenges have limited technical capacity, including limited staff resources with technical skills to conduct extensive research and work on operational support to field teams at the same time. This is the reason why UNHCR Innovation resorted to establishing collaborations with well-known technical experts in the areas of big data analytics and artificial intelligence, such as the United Nations Global Pulse Data Fellows programme and the Human Rights, Big Data and Technology (HRBDT) initiative at the University of Essex to work on Project Jetson. In addition to this, the team hired data visualization and design consultants to be able to visualize the different models, techniques and approaches developed during the evolution of the project, since its inception.

Selecting the variables to understand the drivers of displacement in Somalia: A human rights approach to computational social science

The Project Jetson team started with an exploratory data analysis to try to understand correlations among variables. First, they started by collecting six groups of data from the 18 regions of Somalia: (a) historical data on river levels and (b) rainfall anomalies – both used as proxies for drought affectation on population movement; (c) historical data on violent conflict, such as number of fatalities per region; (d) number of violent incidents per region, as per the ACLED (2022) methodology; and historical data on forcibly displaced population movement, particularly (e) arrivals and (f) departures. The amount of data acquired corresponds to seven years of historical data (2010–2017), processed in a time series analysis by month format. Project Jetson used this combination of data sets under

the assumption that past population movement, climate/weather anomalies and violent conflict could explain people's movement patterns and that these factors were exacerbating displacement in Somalia.

However, before applying statistical and artificial intelligence-based quantitative modelling, UNHCR Innovation Service conducted a mission to Dollo Ado to design the Project, based on a better understanding of the context and human-centred approach. Following a human rights-centred design, the team wanted to (a) cross-validate ground truth data and (b) debunk or validate underlying assumptions in the data, both needed in computational social science. The technical team thought that if they were able to understand decision-making during the flight, they could observe outliers in the data sets.

To do this in computational social science, there should be a recognition that data is human, and behind a number there is a story about fleeing. The team focused on the agency of the people behind those stories and wanted to understand the reasons for fleeing and the journey, as elements to prediction. The Jetson team resorted then to the use of traditional qualitative research methods, like consent-based semi-structured interviews and focus groups conducted in the five different camps in Dollo Ado, to understand motivations to flee in different regions and understand the conditions of displacement of recent refugee arrivals in Ethiopia. In combination with data science techniques, traditional research methods that are aligned with data protection and responsibility principles can strengthen human rights practices in the appropriate design of data science-based, artificial intelligence-based and algorithm-based interventions in humanitarian settings.

In one interview, a refugee stated that prior to fleeing, once severely menaced by extreme groups demanding a portion of their cattle and crops, they needed to sell their goats and transfer their major financial assets to generate cash for their fleeing journey (UNHCR Innovation Service, 2019a). While goats are not a driver of movement per se, as they do not directly affect movement, they are an indication of movement, directly related to the decision to flee. Analysis of seven years of market prices data found there were different correlations in the Somali regions observed, and that there often was a drop in market prices due to excess supply within the neighbouring regions where people were fleeing to. Thanks to these insights, the Jetson team incorporated two additional variables into the data set: historical prices of commodities, such as local goat prices and water drum prices. Project Jetson then is also the first example of behavioral analytics using artificial intelligence in UNHCR, with a human-centred approach.

Preparing and training the data with machine-learning models: Ethics in artificial intelligence modelling

Once new elements and variables that explained displacement were validated by the people the Project is intended to serve, data preparation – meaning the extraction, cleaning and annotation phase – was the biggest component of the Project. This included mining data by using scripts, scraping information from graphs, reports in PDF format, spreadsheets and websites – as some of the data sets were not in a legally open or technically open format for machine reading. The Jetson team compiled 107 months of data for the 18 regions of Somalia, plus 1 cross-border region (Dollo). These were historical data, so the forms of machine learning open source algorithms made use of the temporal dimension of said data.

The data sets or variables used to perform the training of the machine-learning models were monthly aggregated data sets that reflected the following:

(a) Departures and arrivals per region
(b) Market prices per region: local goat prices and water drum prices – sensitive to drought
(c) Weather-related data per region
(d) Violent conflict and fatalities per region

In order to avoid optimization criteria bias or generalization errors, or any other "blind spots" in building the artificial intelligence-based models, as the Massachusetts Institute of Technology (2019) would define, the inputs for several machine-learning models tested were built as a multivariate time series vector. This means the inputs were the number of arrivals per region as the dependent variable, and the other data sets as independent variables. Initially, a total of 90 per cent of the historical data set was used for training and 10 per cent for testing. The validation set was represented by the actual number of arrivals as reported by UNHCR colleagues as they started getting new monthly data, then the models were refit and tested again manually, creating a feedback loop. Thus, the evaluation metrics used to score model accuracies reflected only the validation set, scored against the prediction outputs of the models. The models and feedback loop were always supervised and selected by a human to be able to understand differences and explain results. The Project Jetson team took a human-in-command ethical approach to artificial intelligence-based modelling (European Union, 2019), which means the machine did not select the number of arrivals automatically, even if it represented at the beginning more manual work for the team.

The Project Jetson team also tested other approaches prior to successfully applying multivariate time series analysis using artificial intelligence. The first experiments were focused on statistical regression analysis, and during the trial-and-error phase of artificial intelligence techniques, convolutional neural networks were tested without success. It was not until the commodity market prices were introduced to the model that the predictions started to yield good results. It took months to refine the scope before achieving good performance metrics scores, while manually refining models until they were able to have good explicability elements to avoid the so-called black box approach (Coppi et al., 2021). The scope was narrowed to encompass only one region, and several points had to be removed from the data considered in order to make the model's prediction even moderately accurate (UNHCR Innovation Service, 2019b). Subsequently, the Project Jetson model was able to correctly predict at least one month in advance arrivals in 13 out of the 18 regions of Somalia, with 80 per cent prediction accuracy and small error rates. The remaining five regions had less accurate predictions, at approximately 50 to 65 per cent, which could be attributed to lack of continuity in the data.

In an effort to understand the needs of field staff, while accurately predicting the number of arrivals in each region, UNHCR Innovation Service executed two sequential experiments.

> *Experiment 1*: Predictions of arrivals one month in advance. This methodology proved to be more demanding, since predicting the number of arrivals one month in advance was highly dependent on data on all of the variables being collected in due time, in order for the model to return output predictions for the next month. The modelling technique applied for this experiment was based on genetic and evolutionary algorithms with the use of the tool Eureqa.

Experiment 2: Prediction of arrivals three months in advance. This methodology, operating with a lag of three months, explored all the options of machine-learning algorithms that could be used for time series predictions, such as naive models, classic regression models and neural network models.

The performance of predictions did not decline significantly by changing the forecasting horizon from one month to three months. Once the performance metrics were improving, the team resorted to semi-automating the feedback loop, only for those that yield good performance metrics.

Project Jetson outcomes

Given Project Jetson's experimental nature, no decisions were taken at the operational level; this means no decision was ever done directly using Jetson quantitative predictions by the operation due to initial ethical concerns. These ethical concerns stemmed from the fact that neither UNHCR nor the United Nations System in general has an official guidance or policy regarding artificial intelligence-based or algorithmic-produced predictions to (a) plan/prepare and (b) respond to crises based on algorithmic accountability. Algorithmic accountability examines the process of assigning responsibility for harm when algorithmic decision-making results in discriminatory and inequitable outcomes (Caplan et al., 2018), which currently still lies with human beings and institutions, according to official guidance on accountability to affected populations (UNHCR, 2020). As Hugo Slim (2015) stated, "Difficult humanitarian situations usually take shape gradually, changing and developing over time." Therefore, we argue that using predictive analytics or any artificial intelligence-based system in humanitarian response requires deep understanding of human rights-based approaches and expanding on the professionalization and the study of humanitarian decision-making and humanitarian deliberation, or at least attempting to quantify and evaluate them. Slim (2015) also states that "decision-making is important to the deliberation process as it helps humanitarian agencies make the best possible decisions within their respective organizations."

Ensuring algorithmic accountability, using a human rights approach to the project with humans centred in the design, and focusing on an ethical approach to artificial intelligence-based modelling were fundamental elements in the success of Project Jetson, as an experiment. The inclusion of refugees and IDPs in the design contributed to Jetson's success. Understanding humanitarian decision-making processes during emergency situations (e.g. breaking points in scenarios, thresholds for contingency planning scenarios) was extremely important for designing Jetson's predictions and subsequent visualization (UNHCR Innovation Service, n.d.). The Somalia and Dollo Ado teams increased their ongoing communication regarding movement and other elements that could exacerbate displacement, as a result of this process. This in turn showcased the systemic breaking points in scenario planning and how predictions in the form of maps and number ranges (low, medium, high) could help humanitarians avoid those breaking points and informed contingency planning.

Jetson also highlighted the importance of partnerships, which were strategically leveraged to complement or provide expertise that could not be found in UNHCR (Earney and Moreno Jimenez, 2019:106). Project Jetson led to the establishment of a partnership with the United Nations Global Pulse Data Fellows programme for doctoral researchers (Vacarelu, 2019), to increase the predictive time from one month to three months in advance and test different modelling approaches. Jetson

also could not have been done without the refugees and IDPs in Dollo Ado, who provided project feedback and helped design it. Their feedback and stories shaped the scientific and innovation processes undertaken. Further, Jetson depended on UNHCR colleagues and humanitarian partners on the ground collecting data in some of the most remote locations in Somalia. The involvement of partners is essential, especially of non-traditional partners who broaden perspectives in the ideation, design and scale of technology-based solutions. The Project's collaborative nature highlights the importance of partnerships in achieving new scientific outcomes and improving evidence in refugee crises.

Additionally, using design thinking and considering the field teams' user needs while designing Jetson, UNHCR Innovation Service (2019b) learned that a tool is only as useful as the user's understanding of it. For instance, considering colleagues' needs in low-bandwidth environments led to designing an online portal for Jetson with web pages that were not too heavy to run (ibid.). In parallel with the scoping of the product, the Jetson Project team explored which metrics and what modelling could be shared with colleagues. While Jetson was never operationalized, it provided crucial insights that can be used in the future to develop useful tools in collaboration with users.

Jetson also unveiled the potential for international organizations to elaborate policy and guidance on artificial intelligence or algorithmic accountability around predictive analytics initiatives and/or solutions similar to automated decision-making (ADM) systems that can potentially have a negative impact on crisis-affected populations. Jetson proved the need for open data and population flow data sets for strengthening humanitarian service delivery, preparedness and response. However, whether to build or use automated (or semi-automated) decision-making systems to assist service delivery, preparedness and response efforts presents a main ethical question among humanitarian actors.

Challenges and learnings from Project Jetson

The use of ADM systems, a class of technologies that assist or replace the judgment of human decision makers, poses potentially grave risks when used for border management and/or internal displacement-related humanitarian activities. ADM systems are likely to have important human rights implications regardless of whether they operate autonomously or their outputs are simply one factor considered by a human in rendering a final decision (Molnar and Gill, 2018). Every adoption of nascent technologies like ADM systems poses challenges as they operate without direct human supervision (Kuziemski and Misuraca, 2020). Whenever technology is used on humans, especially technologies that use automation or machine learning, it is imperative that a strong safety net is created to ensure that the technology is not used for harmful purposes (ICRC and Privacy International, 2018). These harmful purposes include surveilling, targeting, or leading to the discrimination and exclusion of individuals in vulnerable situations, such as refugees (OHCHR, 2020). Because Project Jetson represented an avenue for potential border closures, automating contingency planning figures and even potentially targeting of vulnerable people, it was purposefully never used in decision-making, and its data was carefully curated – to the point that only a historical data sample represented the public sites for showcasing purposes.

Additionally, simply having good quality historical population or other type data does not ensure a successful predictive analytics project, as this also requires collaboration with partners in establishing data-sharing agreements as well as periodicity and continuity of data. First, data processing needs to be done by applying strong data protection and data responsibility practices, including anonymization and aggregation of data that represent vulnerable populations who could, for example, be exposed by triangulation with other data sets (ICRC, 2020). Second, there must be a data-sharing agreement for external partners that establishes the conditions for data acquisition, transfer, storage, processing and deletion. Each partner has different requirements, and those need to be stated explicitly in the sharing agreement or memorandum of understanding, as well as the technology requirements needed for transfer, storage or processing (e.g. repository, on-premises or other secure cloud-based solutions with the same diplomatic immunity as physical servers in an office). Finally, the publication of results needs to be consistent with ethical standards and humanitarian principles. All publicly available data need to abide by the principle of "do no harm", thus ensuring that publishing predictions of population movement will not negatively affect the access to territory or human rights of displaced individuals.

This is one of the main reasons why some of the data sets and the specific code for Project Jetson are not publicly available. In order to avoid contradicting the "do no harm" principle, UNHCR Innovation Service worked to ensure that data from Project Jetson were not used for any of the following: (a) human rights exclusions and discriminatory practices, (b) doing harm with technology, and (c) potential misuse of data and predictions for other non-humanitarian purposes. These concerns represent the limitations of predictive analytics projects and their potential for harm, in the absence of adequate human rights safeguards.

Policy implications

Rapidly developing technology has changed the ways in which humanitarians deliver aid to the people they serve. It is imperative that any decision to use automated technology in forced displacement contexts has a solid foundation of ethical and human rights considerations, ensuring that the "do no harm" principle is applied in physical and digital spaces. Jetson was neither intended nor designed as an ADM system to replace humanitarian decision-making, but rather to support it and strengthen humanitarian accountability. Internal UNHCR guidance such as the *Guidance on the Protection of Personal Data of Persons of Concern* (UNHCR, 2018) provides UNHCR staff with important information on how to guarantee privacy and data protection for refugees and IDPs. Data used must be ethically gathered (Slim, 2015) and supported by data protection principles (Hayes, 2017). Project Jetson needed to identify non-traditional variables in order to create models that were able to accurately predict arrival windows. Although many data sets were available, not all were used due to considerations of consistency, relevance and safety (UNHCR Innovation Service, 2019b).

Predictive analytics projects, such as Project Jetson, have enormous potential in the humanitarian sector as they may help understand and anticipate population movements, creating the conditions for anticipatory humanitarian action. Anticipatory humanitarian action shifts action to before a crisis, anticipating the effects and designing mitigation measures to lessen the impact on people in vulnerable situations (Pichon, 2019). This may allow humanitarians to reference predictive models, design action plans or even prearrange financing before an emergency (Centre Team, 2020).

However, predictive analytics is not a panacea to the challenges of understanding displacement and population movement. Before designing a predictive analytics solution, policymakers hoping to better understand forced displacement must identify a clear challenge question and set attainable expectations about what the data can be used to predict and understand (Hofman et al., 2017). Additionally, all displacement crises are unique, shaped by different variables and conditions. When thinking about applying predictive analytics or scaling mathematical or algorithm-based models or indexes, it is important to determine what the thresholds are and what will happen when they are crossed; failing to do so will not allow the use of the data to inform humanitarian response. When deciding if one should build any predictive system, it is imperative to consider and co-design with end users – in this case, displaced individuals and host communities. Failing to place displaced people at the centre of any conversation about whether a technology should be deployed will fundamentally be at odds with humanitarian accountability, as it is those affected that should shape "when and how emerging technologies should be integrated into refugee camps, border security or refugee hearings – if at all" (Molnar, 2019:9).

As humanitarians increasingly turn to technological solutions to better anticipate and respond to the needs of affected populations, it becomes imperative to do no harm in the digital space. Using automated technologies requires creating strong human rights and digital protection systems, consulting affected populations and building robust safeguards against the misappropriation of data/tools by bad actors or States. Humanitarian organizations must balance the desire for data innovation with mitigating or perhaps altogether avoiding the risks posed by those very technologies.

REFERENCES*

Armed Conflict Location and Event Data Project (ACLED)

2022 Quick guide to ACLED data. Available at https://acleddata.com/acleddatanew//wp-content/uploads/dlm_uploads/2022/06/ACLED_GeneralUserGuide_June2022.pdf.

Caplan, R., J. Donovan, L. Hanson and J. Matthews

2018 *Algorithmic Accountability: A Primer*. Data & Society. Available at https://datasociety.net/library/algorithmic-accountability-a-primer/.

Centre Team

2020 Anticipatory action in Bangladesh before peak monsoon flooding. Centre for Humanitarian Data. Available at https://centre.humdata.org/anticipatory-action-in-bangladesh-before-peak-monsoon-flooding/.

Coppi, G., R. Moreno Jimenez and S. Kyriazi

2021 Explicability of humanitarian AI: A matter of principles. *Journal of Humanitarian Action*, 6:19.

Earney, C. and R. Moreno Jimenez

2019 Pioneering predictive analytics for decision-making in forced displacement contexts. In: *Guide to Mobile Data Analytics in Refugee Scenarios: The "Data for Refugees Challenge" Study* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 101–119.

European Union

2019 *Ethics Guidelines for Trustworthy AI.* High-Level Expert Group on Artificial Intelligence, Brussels. Available at https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html.

Hayes, B.

2017 Migration and data protection: Doing no harm in an age of mass displacement, mass surveillance and "big data". *International Review of the Red Cross: Migration and Displacement*, 99(1):179–209.

Hofman, J.M., A. Sharma and D.J. Watts

2017 Prediction and explanation in social systems. *Science*, 355(6324):486–488.

** All hyperlinks were working at the time of writing this report.*

Hövelmann, S.

2020 Humanitarian topics explained: Triple Nexus to go. Centre for Humanitarian Action (CHA), Berlin. Available at www.chaberlin.org//wp-content/uploads/2020/03/2020-03-triple-nexus-to-go-hoevelmann-en-online.pdf.

International Committee of the Red Cross (ICRC)

2020 *Handbook on Data Protection in Humanitarian Action*. Second edition. Available at www.icrc.org/en/publication/430501-handbook-data-protection-humanitarian-action-second-edition.

ICRC and Privacy International

2018 *The Humanitarian Metadata Problem: "Doing No Harm" in the Digital Era*. Available at www.icrc.org/en/download/file/85089/the_humanitarian_metadata_problem_-_icrc_and_privacy_international.pdf.

Kuziemski, M. and G. Misuraca

2020 AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, 44(6):101976. Available at https://doi.org/10.1016/j.telpol.2020.101976.

Massachusetts Institute of Technology (MIT)

2019 AI Blindspot: A discovery process for spotting unconscious biases and structural inequalities in AI systems. Available at https://aiblindspot.media.mit.edu/.

Molnar, P.

2019 New technologies in migration: Human rights impacts. *Forced Migration Review*, 61:7–9. Available at https://search.proquest.com/openview/c663a48a99e46dfb6dc70d3584b1cf2e/1?pq-origsite=gscholar&cbl=55113.

Molnar, P. and L. Gill

2018 *Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System*. The Citizen Lab and the International Human Rights Program, University of Toronto Faculty of Law. Available at https://citizenlab.ca/wp-content/uploads/2018/09/IHRP-Automated-Systems-Report-Web-V2.pdf.

New York University Tandon School of Engineering (NYU Tandon)

2017 The GovLab at NYU Tandon launches website on "data collaboratives" – New forms of public–private data exchanges that create public value. Press release. 17 January. Available at https://engineering.nyu.edu/news/govlab-nyu-tandon-launches-website-data-collaboratives-new-forms-public-private-data-exchanges.

Office of the United Nations High Commissioner for Human Rights (OHCHR)

2020 Racial and xenophobic discrimination, emerging digital technologies in border and immigration enforcement. Available at https://undocs.org/A/75/590.

Office of the United Nations High Commissioner for Refugees (UNHCR)

2014 Information Management Toolkit. Geneva. Available at https://im.unhcr.org/imtoolkit/.

2018 *Guidance on the Protection of Personal Data of Persons of Concern to UNHCR*. Geneva. Available at www.refworld.org/docid/5b360f4d4.html.

2020 *Operational Guidance on Accountability to Affected Populations (AAP)*. Available at www.unhcr.org/handbooks/aap/.

UNHCR Innovation Service

2019a A goat story. 9 May. Available at https://medium.com/unhcr-innovation-service/a-goat-story-3ed6bdd2b237.

2019b A recipe for success: Humility and resilience in humanitarian innovation failures. 14 May. Available at https://medium.com/unhcr-innovation-service/the-recipe-for-success-humility-and-resilience-in-humanitarian-innovation-failures-d8d72276ceaa.

n.d. Project Jetson. Available at http://jetson.unhcr.org.

Pichon, F.

2019 Anticipatory humanitarian action: What role for the CERF? Working paper No. 551. ODI, London. Available at https://cerf.un.org/sites/default/files/resources/ODI_Early_Action_Study.pdf.

Protection and Return Monitoring Network (PRMN)

2017 Notes on methodology. UNHCR Somalia. Available at https://data2.unhcr.org/en/documents/details/53888.

n.d. UNHCR Somalia internal displacement. UNHCR and Norwegian Refugee Council (NRC). Available at https://unhcr.github.io/dataviz-somalia-prmn/.

Seal, A. and R. Bailey

2013 The 2011 famine in Somalia: Lessons learnt from a failed response? *Conflict and Health*, 7:22.

Slim, H.

2015 Ethical practice in humanitarian action – Humanitarian deliberation. In: *Humanitarian Ethics: A Guide to the Morality of Aid in War and Disaster*. Hurst & Company, London. Available at https://cadmus.eui.eu/handle/1814/63625.

Vacarelu, F.

2019 Global Pulse launches Data Fellows programme to connect doctoral researchers with UN entities. United Nations Global Pulse, 3 February. Available at www.unglobalpulse.org/2019/02/global-pulse-launches-data-fellows-programme-to-connect-doctoral-researchers-with-un-entities/.

ADDITIONAL READING

Centre Team

2019 Workshop recap: Predictive analytics in humanitarian response. Centre for Humanitarian Data. Available at https://centre.humdata.org/workshop-recap-predictive-analytics-in-humanitarian-response/.

German Federal Foreign Office (FFO) and International Organization for Migration (IOM)

2019 *Workshop Report on Forecasting Human Mobility in Contexts of Crises*. Displacement Tracking Matrix (DTM). Available at www.alnap.org/help-library/workshop-report-on-forecasting-human-mobility-in-contexts-of-crises-displacement.

Nair, R.

2019 Machine learning in action for the humanitarian sector. IBM Research – Ireland, 21 January. Available at www.ibm.com/blogs/research/2019/01/machine-learning-humanitarian-sector/.

Office for the Coordination of Humanitarian Affairs (OCHA)

2020 *Peer Review Framework for Predictive Analytics in Humanitarian Response*. The Centre for Humanitarian Data. Available at https://centre.humdata.org/predictive-analytics/.

Open Knowledge Foundation

n.d. The Open Data Handbook. Available at http://opendatahandbook.org/guide/en/.

Pasquale, F.

2019 The second wave of algorithmic accountability. LPE Project. Available at https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/.

Powles, J. and H. Nissenbaum

2018 The seductive diversion of "solving" bias in artificial intelligence. *OneZero*, 8 December. Available at https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53.

Save the Children International

2018 *Predicting Displacement: Using Predictive Analytics to Build a Better Future for Displaced Children*. London. Available at https://resourcecentre.savethechildren.net/node/14290/pdf/predicting_displacement_report_-_save_the_children_mdi.pdf.

Suleimenova, D., D. Bell and D. Groen

2017 A generalized simulation development approach for predicting refugee destinations. *Scientific Reports*, 7(13377).

United Nations

2011 UN declares famine in two regions of southern Somalia. UN News, 20 July. Available at https://news.un.org/en/story/2011/07/382072-un-declares-famine-two-regions-southern-somalia.

UNHCR

1951 Convention relating to the status of refugees. Geneva, 28 July. Available at www.unhcr.org/5d9ed32b4.

1967 Protocol relating to the status of refugees. New York, 31 January. Available at www.unhcr.org/5d9ed66a4.

2011 Tens of thousands of drought-displaced Somalis head to Mogadishu. 26 July. Available at www.unhcr.org/news/latest/2011/7/4e2eca339/tens-thousands-drought-displaced-somalis-head-mogadishu.html.

2015 *Policy on the Protection of Personal Data of Persons of Concern to UNHCR*. Available at www.refworld.org/pdfid/55643c1d4.pdf.

n.d. SAM Project – Results of the model. Available at https://chao-unhcr.shinyapps.io/SAMshiny/.

UNHCR Innovation Service

2016 Migration, mitigation and maps: The predictive role of UNHCR's first Winter Cell. Available at www.unhcr.org/innovation/migration-mitigation-and-maps-the-predictive-role-of-unhcrs-first-winter-cell/.

United Nations System Chief Executives Board for Coordination (CEB)

2019 *Report of the High-level Committee on Programmes at Its Thirty-eighth Session*. International Training Centre of the International Labour Organization, Turin, Italy, 10 and 11 October 2019. Available at https://undocs.org/CEB/2019/6.

BRIDGING SURVEY-BASED ESTIMATES AND AIRLINE PASSENGER DATA TO PRODUCE PUERTO RICO NET MIGRATION ESTIMATES IN THE AFTERMATH OF HURRICANE MARIA

Jason Schachter,[1] *Angelica Menchaca*[2] *and Antonio Bruce*[3]

Introduction

Whether measuring migration stocks, flows or net rates, traditional methods of measuring migration and sources of migration data are often lacking for various reasons. While each data source has its purpose and distinct advantages, limitations often hinder functionality. By adapting their approach, social scientists can potentially improve the accuracy and timeliness of migration statistics by incorporating big data into their methods. "Big data" can be defined as copious amounts of digital information that can be processed, stored and analysed to produce migration statistics (Ashton et al., 2016). Big data are currently not seen as a replacement for traditional sources of international migration data, but rather as a way to supplement and add value to pre-existing data sources (IOM, 2017). There has been a fair amount of research into the application of big data to measure international migration in recent years (for examples, see: IOM, 2021). One such example is the United States Census Bureau's recent use of airline passenger data, combined with the American Community Survey (ACS), to measure the impact of Hurricane Maria on migration out of Puerto Rico.

Investigating supplementary data sources to better reflect net migration estimates after Hurricane Maria

In September 2017, Category 5 Hurricane Maria made landfall on the island of Puerto Rico, resulting in extensive damage, loss of human life, and outmigration to mainland United States. The Commonwealth of Puerto Rico is an unincorporated territory of

[1] Jason Schachter is Chief of the United States Census Bureau's International Migration Branch. He previously worked as a statistician for the United Nations Economic Commission for Europe and as Population Affairs Officer at the United Nations. He also held the position of Director of Research for New York City's Division of Citywide Equal Employment Opportunity and was an Adjunct Professor at Columbia University's School of International and Public Affairs. He holds a PhD in Rural Sociology/Demography from the Pennsylvania State University.

[2] Angelica Menchaca is a survey statistician/demographer for the Population Division's International Migration Branch of the United States Census Bureau. Angelica received a doctorate in Sociology with a focus on International Migration and Demography from Texas A&M University. Her focus is on demographic data analysis and statistical modelling using administrative and survey-based data sources.

[3] Antonio Bruce is a mathematical statistician with the Population Evaluation, Analysis and Projections Branch of the United States Census Bureau. Antonio received a master's in Applied Statistics with a focus on Multivariate Data Analysis from the George Washington University. His focus is on demographic data analysis and multivariate statistical modelling using administrative and survey-based data sources.

the United States of America, with a population of over 3 million people. Puerto Rico's population has been declining since 2004, primarily due to outmigration to the United States, coinciding with economic decline on the island (Puerto Rico Report, 2013). As United States citizens, with the right to free movement between the Commonwealth and the mainland, over 5 million Puerto Ricans are living in the United States (Wang and Rayer, 2018). Natural disasters can impact the population – namely, through the movement of people from affected areas, as well as through deaths resulting from cataclysmic events. As natural disasters increase in frequency and magnitude (Smith, 2019; WMO, 2019), so do the needs for population estimates programmes to accurately measure their impact, often requiring different data sources or the implementation of new methods. Previous research was conducted by the Census Bureau, evaluating the displacement and relocation of victims post-Hurricane Katrina, using ACS and Federal Emergency Management Agency (FEMA) data (Koerber, 2006).

Initial reports varied as to the size of the potential exodus to the United States. The Governor of Puerto Rico predicted millions of Puerto Ricans moving to the United States, while other estimates ranged from 100,000 to 240,000 people (for example, see: Meléndez and Hinojosa, 2017). Between 3 October and 30 November, the Florida Division of Emergency Management counted over 208,000 people from Puerto Rico landing in Miami airports. However, these early estimates did not account for potential return migration to Puerto Rico, which was estimated to range from 135,000 to 145,000, resulting in a net outmigration of 90,000 to 93,000 people (Krogstad, 2015; Rayer, 2018).

Household sample surveys like the ACS are not designed to pick up sudden mass movements of people, since retrospective survey-based migration data tend to lag behind actual migration events. Surveys do not measure a migration event in real time, but rather measure the migration event at the time when a survey respondent is included in the sample. This works well when migration patterns are stable, but when there are large annual fluctuations in the magnitude of movement, those will not be fully picked up until later (usually in the following survey year). The late timing of Hurricane Maria (late September) in the survey data collection cycle, and the corresponding short period of time left to include those affected in the sample, created potential response complications. As a result, it was necessary to look for an alternative data source to be able to measure the impact of Hurricane Maria on migration out of Puerto Rico.

The former methodology used to estimate migration between Puerto Rico and the United States utilized the ACS and its counterpart, the Puerto Rico Community Survey (PRCS). As a result of the hurricane, ACS and PRCS migration flow estimates yielded inadequate measures of net migration for the 2018 estimates year. However, monthly airline data did show a mass exodus of people from Puerto Rico to the United States during this time period. In our new methodology, ACS/PRCS data were integrated with monthly airline passenger traffic (APT) data from the Bureau of Transportation Statistics (BTS) to improve previous estimates. Historically, APT data have consistently shown higher net outmigration from Puerto Rican to the United States than ACS/PRCS estimates. To account for this inherent difference between data sources, the revised method blends ACS/PRCS and APT data. The result is an estimate of the Puerto Rican population impacted by Hurricane Maria.

Traditional data sources: American Community Survey and Puerto Rico Community Survey

The Census Bureau annually produces population estimates representing the population as of 1 July of each year. The ACS and PRCS are annual, continuous household surveys of the United States population that ask detailed information previously collected on the decennial census long-form questionnaire. The ACS currently surveys about 3.5 million households per year, while the PRCS has a sample of 36,000 Puerto Rican addresses. Using the previous methodology, estimates of migration flows from Puerto Rico to the United States were based on responses to the residence one year ago (ROYA) question, which asks where respondents lived one year prior to the survey. Data are collected on a continuous basis throughout the calendar year, though movement could have occurred at any time over a two-year period, depending on when the respondent was included in the sample and when they actually moved. Net migration from Puerto Rico impacts the population estimate not only for Puerto Rico (which can change only via births, deaths or migration), but also for mainland United States, as it is included as part of the mainland's net international migration component.

The 2016 ACS and PRCS measured 88,000 migrants from Puerto Rico to the United States, and 21,000 migrants from the United States to Puerto Rico, resulting in a net loss of 67,000 people for Puerto Rico. The results in 2017 were similar, with 97,000 migrants from Puerto Rico and 20,000 migrants to Puerto Rico, resulting in a net loss of 77,000. The change of just 10,000 net migrants between 2016 and 2017 indicates that the 2017 ACS did not reflect much hurricane-related movement to the United States, at least not to the magnitude expected.

Air passenger traffic data as an alternative data source

In the United States, flight data are compiled from monthly reports filed by over 200 commercial United States and foreign air carriers with the BTS, including both domestic and international flights. Release of annual data on international flights lags about six months from the end of the year, with complete international data available in the middle of June the following year. Domestic flight data are released more frequently (lagging about three months). Monthly international flight data are released three months after domestic data. Reporting of data is for all flights (thus no sampling is involved), following federal report guidelines which went into effect in October 2002.

For Puerto Rico, APT domestic data provide monthly information on the number of passengers flying on planes between Puerto Rico and mainland United States. It should be noted that APT data include information on all travellers, without differentiation of passenger type, thus also including tourists and visitors, who make up the majority of passengers. Non-migrants are counted on both inbound and outbound flights, while migrants are counted only in one direction – unless leaving temporarily, in which case they would be counted again upon their return. A limitation of this method is that it can provide a number for net migration only, with no information on total inflows or outflows, as migrants cannot be distinguished from the total number of passengers entering or leaving Puerto Rico. Additionally, this method is applicable only to a country or territory without any land borders, such as an island like Puerto Rico, as flights are the predominant method of arriving or leaving.

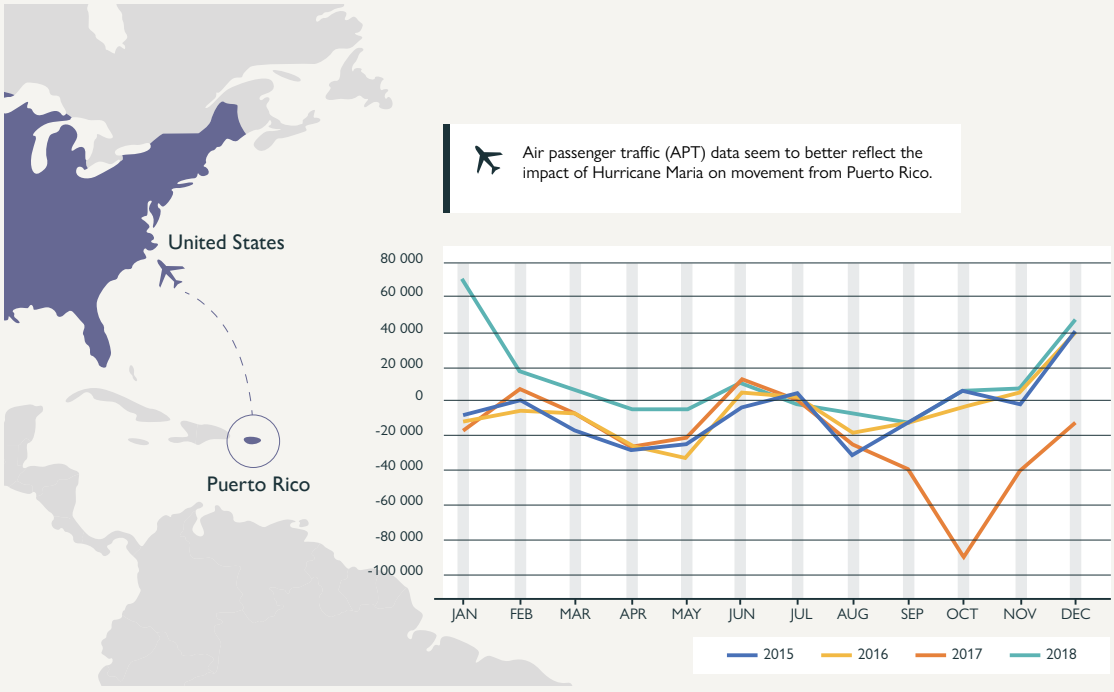
Monthly tallies of net passenger flow movement reflect seasonal variation related to tourism, with presumably greater movement into and out of Puerto Rico in the summer and winter vacation months. Ship passenger movement to and from Puerto Rico is assumed to be minimal. Depending on

the measurement period (e.g. a calendar year), this could lead to year-to-year fluctuations related to annual tourism trends. For example, a high number of tourists could be counted in December, while the return of these same tourists might happen in January of the following year. Over time, these fluctuations are thought to balance out.

Comparing the results of airline passenger traffic data with conventional data sources

APT data seem to better reflect the impact of Hurricane Maria on movement to and from Puerto Rico than ACS/PRCS data alone. Figure 1 shows a large net outmigration in the latter months of 2017 (September to December), corresponding with return migration in the early months of 2018. Looking at 2018 data, this return movement to Puerto Rico from the United States dropped during the early months of 2018, returning to net outmigration by April 2018. Prior to Hurricane Maria (20 September 2017), net movement between the United States and Puerto Rico followed relatively stable monthly patterns, with more passengers leaving than entering Puerto Rico, except for some summer or winter months (June and December, in particular). This corresponds with seasonal flight patterns, with more tourists coming during summer and winter months, as well as return visits during vacation periods by Puerto Ricans living in the United States.

Figure 1. Monthly net flight passenger movement between Puerto Rico and the United States: 2015 to 2018

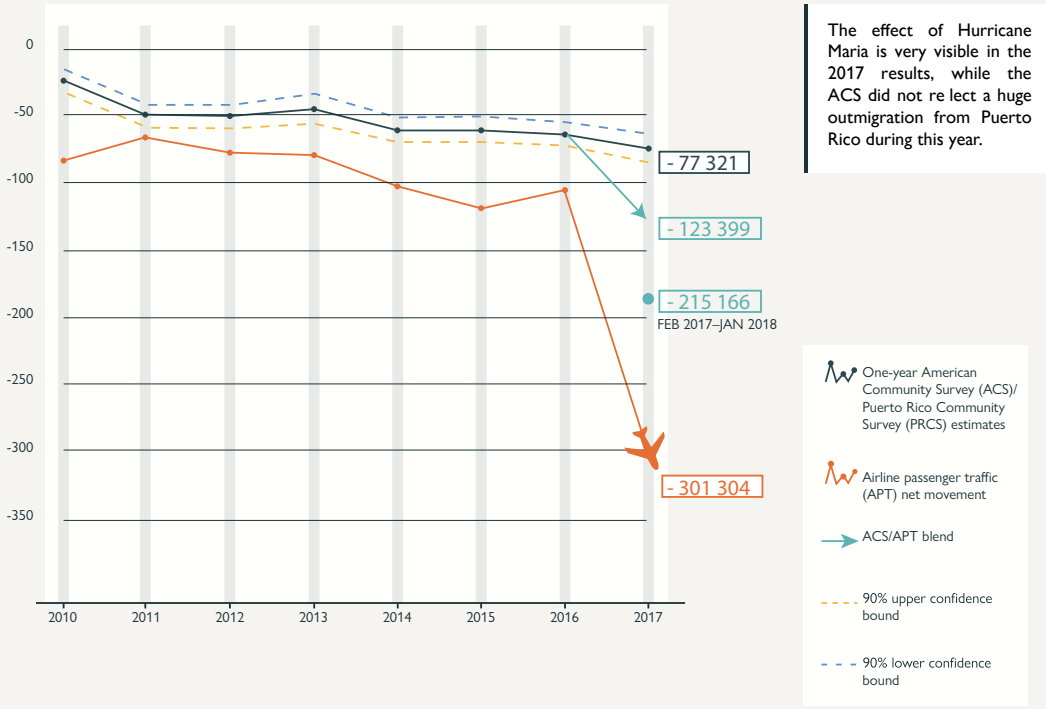


Source: BTS Form 41, T100 (International) Segment All Carriers.

The pattern changes for 2017, as we see the first surge in passengers leaving Puerto Rico in September, peaking in October, and continuing into months when flight passengers typically enter Puerto Rico. The year 2018 is also an anomaly, as we see a large number of returns to Puerto Rico in January and February, and even positive movement in March. Movement normalizes thereafter to previous patterns. It is evident that movement to and from Puerto Rico from September 2017 to February 2018 was particularly impacted by Hurricane Maria, despite 2017 ACS/PRCS data not reflecting this.

As seen in Figure 2, ACS and APT data tended to follow similar patterns prior to 2017, with APT data consistently showing more net out-movement than ACS net outmigration. The effect of Hurricane Maria is very visible in the 2017 results, while the ACS did not reflect a high rate of outmigration from Puerto Rico during this year.

Figure 2. Net Puerto Rico-to-United States migration: 2010 to 2017



Source: Census Bureau, ACS and PRCS; BTS Form 41, T100 (International) Segment All Carriers.

Combining airline traffic data with household sample surveys

Methodology used to combine American Community Survey and airline passenger traffic estimates

To make the two data sources as comparable as possible, we compiled monthly flight data for the 2017 calendar year to coincide with the ACS/PRCS estimation period.[4] We also limited flight information to domestic flights between the United States and Puerto Rico, excluding international flights. The final method applied a simple ratio, using the ratio of ACS/PRCS-to-APT net migration results over a two-year period: 2015 to 2016. We also investigated using ratios based on longer time periods (as far back as 2012) but opted to go with a shorter time period to more accurately reflect recent relationships between data sources. The calculated ratio was applied to the APT Puerto Rico–United States net migration figure measured for calendar year 2017, to remain methodologically consistent with previous ACS/PRCS-based estimates.

[4] This differs from the 2018 estimates year, which covers 1 July 2017 to 30 June 2018.

Final method adjustment

The initial application of the ratio method to 2017 APT calendar year data resulted in a net outmigration of -172,799 people for 2017. Since the Census Bureau's 2018 population estimates represent the population on 1 July 2018, we also considered return migration to Puerto Rico in early 2018. In order to account for return migration in January, we used the 12-month APT time period from February 2017 to January 2018 prior to applying our two-year ratio to make the time period as ACS-equivalent as possible, while still taking into consideration post-Hurricane Maria return migration to Puerto Rico. Shifting our time period one month helped account for return migration yet also kept most months (11 of 12) within the ACS/PRCS-equivalent 2017 calendar year. This modification resulted in an APT Puerto Rico–United States net migration figure of -215,166, which when adjusted by the APT-ACS ratio yielded a final figure of -123,399 net migration between Puerto Rico and the United States. The final time series results are reflected in Figure 2. This method takes APT data from approximately the same time period as the ACS (shifted one month), then adjusts this figure as if it were the ACS based on trends over the past two years (i.e. as if the ACS was able to measure this sudden migration event), while also accounting for return migration early in 2018.

Results

As seen in Table 1 below, ACS/PRCS net migration results were, on average, lower than the APT by a factor of 0.5735 during the two-year period of 2015 and 2016 (pre-Hurricane Maria patterns). Thus, to get an equivalent ACS/PRCS figure for 2017, the 2017 APT number (-215,166) was multiplied by the 0.5735 ratio to yield an estimate of -123,399 people. This is equivalent to applying 2017 ACS/PRCS data to the 2018 estimates year, consistent with the methodology used for previous Puerto Rico-to-United States net migration estimates. Net migration estimates from flight data were twice the size of migration rates indicated by ACS/PRCS data, suggesting that traditional data may underestimate actual movement in times of natural disasters.

Table 1. ACS/PRCS-to-APT ratio method: 2015 to 2017

Year	Net outmigration estimate		Ratio
	ACS/PRCS	APT	
2015	-64 238	-122 084	0.52618
2016	-67 480	-108 693	0.62083
2017	-123 399*	-215 166	0.5735**

* This is the 2017 ACS/PRCS value derived from applying the 2015–2016 average ratio to the 2017 APT total.

** The 2017 ratio is the average of the 2015 and 2016 ratios.

Source: BTS, Air Carrier Statistics (T-100); Census Bureau, ACS and PRCS.

Conclusion

The application of big data is still limited from the perspective of the United States for producing net international migration estimates, but we are exploring alternative sources to improve and benchmark net international migration estimates produced via our current methodology. Given time constraints and limited internally available data, the recency and accessibility of airline passenger data made it ideal for the given circumstance. Accessibility, validity, recency and reliability are fundamental components when exploring alternative data sources, while flight data were particularly applicable for an island like Puerto Rico.

Integration of multiple data sources will play a prominent role in the future production of international migration estimates and will continue to be pursued and incorporated into the Census Bureau's population estimates programme. For example, similar efforts were made to account for the impact of COVID-19 on United States international migration patterns in 2020. Conjoining data sources provided a more reliable estimate for the migration movement resulting from Hurricane Maria than one single source. More research on the validity of flight data, as well as how they are collected and measured, is needed and will be an area for future work. Flight records indicated much larger migration net outflows than survey data, thus highlighting the importance of incorporating benchmark and supplementary data. Relying on one data source alone may limit the ability of researchers to fully capture migration movement.

REFERENCES*

Ashton, W., P. Bhattacharyya, E. Galatsanou, S. Ogoe and L. Wilkinson

2016 Emerging uses of big data in immigration research. Final report submitted to the Social Sciences and Humanities Research Council of Canada.

International Organization for Migration (IOM)

2017 Big data for migration: Uses, opportunities and challenges. United Nations Expert Group Meeting on Improving Migration Data in the Context of the 2030 Agenda. UN Headquarters, New York, 20–22 June. Available at https://unstats.un.org/unsd/demographic-social/meetings/2017/new-york--egm-migration-data/Session%209/Session%209%20IOM.pdf.

2021 Big data, migration and human mobility. Available at https://migrationdataportal.org/themes/big-data-migration-and-human-mobility.

Koerber, K.

2006 Migration patterns and mover characteristics from the 2005 ACS Gulf Coast area special products. United States Census Bureau. Available at www2.census.gov/topics/preparedness/pdf/gulf_migration.pdf.

Krogstad, J.M.

2015 Puerto Ricans leave in record numbers for mainland U.S. Pew Research Center. Available at www.pewresearch.org/fact-tank/2015/10/14/puerto-ricans-leave-in-record-numbers-for-mainland-u-s/.

Meléndez, E. and J. Hinojosa

2017 Estimates of post-Hurricane Maria exodus from Puerto Rico. Research brief. Centre for Puerto Rican Studies. Available at https://centropr.hunter.cuny.edu/sites/default/files/RB2017-01-POST-MARIA%20EXODUS_V3.pdf.

Puerto Rico Report

2013 Puerto Rico's population continues to decline. 7 January. Available at www.puertoricoreport.com/puerto-ricos-population-continues-to-decline/#.XVa1Hf2P4fc.

Rayer, S.

2018 Estimating the migration of Puerto Ricans to Florida using flight passenger data. Bureau of Economic and Business Research, Population Studies. Available at www.bebr.ufl.edu/sites/default/files/Research%20Reports/puerto_rican_migration.pdf.

** All hyperlinks were working at the time of writing this report.*

Smith, A.B.

2019 2018's billion dollar disasters in context. 7 February. Available at www.climate.gov/news-features/blogs/beyond-data/2018s-billion-dollar-disasters-context.

Wang, Y. and S. Rayer

2018 Growth of the Puerto Rican population in Florida and on the U.S. mainland. Bureau of Economic and Business Research, Population Studies. Available at www.bebr.ufl.edu/articles_publication/growth-of-the-puerto-rican-population-in-florida-and-on-the-u-s-mainland/.

World Meteorological Organization (WMO)

2019 *WMO Statement on the State of the Global Climate in 2018*. No. 1233. Geneva. Available at https://library.wmo.int/doc_num.php?explnum_id=5789.


71277
26722
93499
37833
04500
82388
59055
60166
15611

MONITORING PUBLIC SENTIMENTS AND ENGAGING MIGRANT COMMUNITIES

HOW CAN BIG DATA ANALYTICS HELP UNDERSTAND MIGRANT INTEGRATION?

Tuba Bircan,[1] Albert Ali Salah[2] and Alina Sîrbu[3]

Introduction

Migration has always been an integral part of history and will certainly shape the future of societies. Migrant integration and inclusion are complex phenomena influenced by multiple and interlinked factors, and they present both challenges and opportunities for global sustainable development. Recent migration trends highlight the significance of recognizing and tackling questions regarding diversity and social cohesion across different world regions. Successful inclusion of immigrants is a prerequisite for social cohesion and economic progress.

Adequate data are key for evidence-based policymaking. However, while a large amount of official statistics is produced across European Union member States, only a small part of the complexity of migrant integration phenomena can be captured through such data. Social security databases to assess labour market integration allow counting migrants only after they officially enter the systems. On the other hand, indicators to explain the different aspects of migrants' living conditions and monitor integration policies implemented by countries cannot be captured through register data; large-scale surveys such as the European Union Labour Force Survey and the European Union Survey on Income and Living Conditions are the major data sources for these aims. However, migration and migrants have never been the centre of attention (with the exception of some occasional ad hoc modules) for cross-sectional/multinational surveys. These are designed to cover the general resident population in a given country, and although the migrants/non-nationals are incorporated in samples, the representation and coverage of those migrants have been questionable. Consequently, to improve understanding of migrant integration, it is necessary to go beyond traditional data sources, and novel approaches are required.

[1] Tuba Bircan is Research Professor in the Department of Sociology and Research Coordinator of Interface Demography at the Free University of Brussels. She is also affiliated as a Senior Research Associate with the Higher Institute for Labour Studies, Catholic University of Leuven. She is the corresponding author and can be contacted at tuba.bircan@vub.be and Tuba.Bircan@kuleuven.be.

[2] Albert Ali Salah is Professor and Chair of Social and Affective Computing at the Department of Information and Computing Sciences of Utrecht University, the Netherlands. He is also affiliated with the Department of Computer Engineering of Boğaziçi University, Türkiye. He can be contacted at a.a.salah@uu.nl.

[3] Alina Sîrbu is Associate Professor of Computer Science at the University of Pisa, Italy. She is a member of the Knowledge Discovery and Data Mining Laboratory and the Computational Health Group of her department. She holds a PhD in Computer Science from Dublin City University, Ireland. She can be contacted at alina.sirbu@unipi.it.

Understanding the various dimensions of the governmental and public commitment to migrant integration is essential to design targeted interventions and, ultimately, enable migrants and refugees to fully engage with society. For this purpose, useful indicators should be derived from large data repositories. Several past efforts used data science and machine-learning techniques to study migrant integration via different indicators. To give more specific examples, global bilateral remittance flows were mapped using big data visualization techniques for the World Banks's 2011 bilateral remittance database, which estimates flows across 215 countries. Another experimental approach was implemented by PeaceTech Lab, where social media data from South Sudan were analysed and machine-learning-based visualizations for Myanmar were used to better monitor hate speech and public sentiment (Monroe, 2018). An Italian study analysed retail data from a supermarket chain and observed immigrants' food consumption baskets to understand to what extent immigrants converge with or diverge from the norms and habits of the destination country (Sîrbu et al., 2021). Other examples include, but are not limited to, the analysis of localized language use and diversity in Twitter data using half a billion geotagged tweets (Magdy et al., 2014), and the analysis of a data set combining mobile phone data with media events data and data on housing market transactions in Türkiye to determine refugee settlement locations (Bansak et al., 2018). This chapter will contribute to the existing efforts by elaborating two special cases, namely the use of social media data and mobile phone data for studying various aspects of migrant and refugee integration.

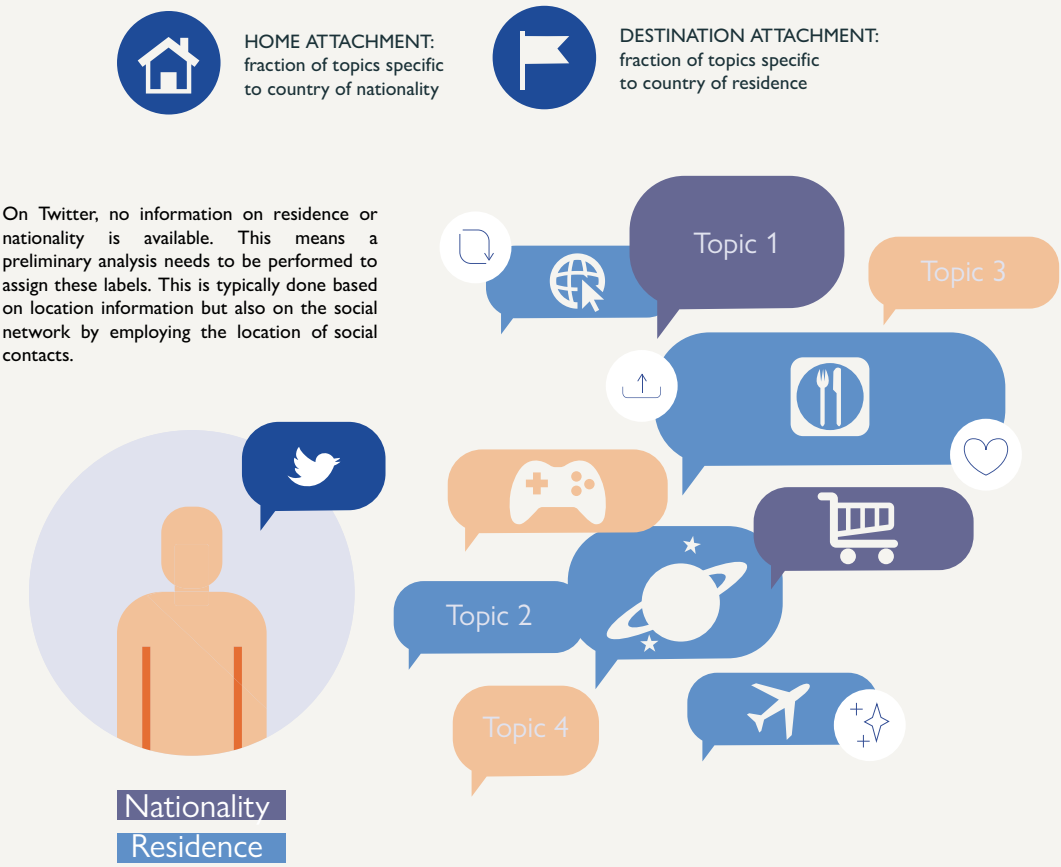
Case studies

Measuring acculturation through Twitter data analysis

The Twitter platform hosts millions of users exchanging daily information in the form of short messages. Besides the messages themselves, the platform also contains the social network of the users. Much of this information is public and can be downloaded using the public Twitter application programming interfaces (APIs), resulting in data that can be employed to study migrant acculturation. Acculturation can be defined as "the process of group and individual changes in culture and behaviour that result from intercultural contact" (Berry, 2019), which is always present in the case of human migration. Acculturation does not merely imply that a culture is absorbed by another, but it is more of a continuous and bilateral process where cultures evolve through contact. We can thus observe different acculturation patterns in different individuals, and Twitter can be employed to study these, as outlined below.

A first step in studying migrant behaviour, including the process of acculturation, is to identify migrants in the data set. While in traditional migration data, this step is intrinsic to the data-collection process, where participants of censuses or surveys are asked to declare their nationality and/or country of birth, along with country of residence, big data from social media and other human digital traces often do not include this information. This is also the case with Twitter, where no information on residence or nationality is available. This means that a preliminary analysis needs to be performed to assign these labels. This is typically done based on location information but also on the social network, using the location of social contacts (Kim et al., 2020). With the nationality and residence labels assigned, migrants can be defined as users who have a country of residence different from their country of nationality (or birth).

Figure 1. Computing home attachment and destination attachment for Twitter users

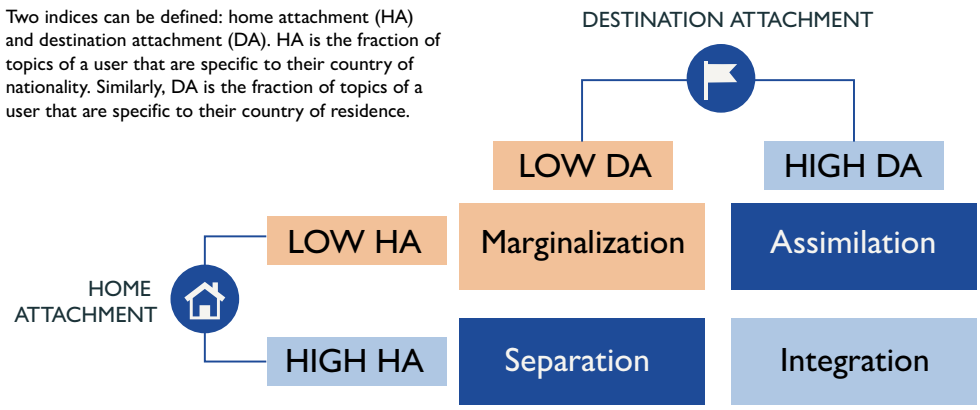


Note: The users are assigned a nationality and residence. Then the topics they discuss are studied. Topics specific to the country of residence and nationality are used to compute the home attachment and destination attachment indices.

Once migrant users are selected, acculturation can be measured in various ways. One possible approach is studying the topics that they discuss. If migrants are observed to be involved in conversations specific to their country of origin, they are assumed to maintain a cultural link to their home country. Participation in conversations related to their country of residence, on the other hand, is a sign of absorbing the culture of the destination country. We can identify topics of conversation through the hashtags that users attach to their messages. To each hashtag, we can assign a nationality based on which country has native-born users who post that hashtag most frequently.

The study of the topics discussed provides information on multiple aspects related to the process of acculturation: how much information about a specific country reaches a certain individual; how much an individual is interacting with peers, i.e. social integration; how much an individual is engaged politically and civically, etc. Technically, two indices can be defined: home attachment and destination attachment (Kim et al., 2021), as displayed in Figure 1. Home attachment is the fraction of topics of a user that are specific to their country of nationality. Similarly, destination attachment is the fraction of topics of a user that are specific to their country of residence. Identification of topics on Twitter is easily performed by extracting hashtags, while the country of the topic can be assigned by looking at the nationality of the non-migrant users tweeting with those hashtags.

Figure 2. Conceptualizing the four types of acculturation using home attachment (HA) and destination attachment (DA)



The concepts of home attachment and destination attachment help identify the type of acculturation process that migrants are undergoing (Berry, 2019) by considering specific regions in the space defined by the two dimensions, as outlined in Figure 2. *Separation* is assumed when destination attachment is low and home attachment is high for those persons who maintain strong links with their home country and do not appear to build links to their destination. At the opposite side, we have *assimilation*, with high destination attachment and low home attachment (i.e. those individuals who are immersed completely in the destination culture and lose contact with their origins). Two other known acculturation types are possible. *Integration* applies to high destination attachment and high home attachment (i.e. those individuals who maintain links to their home country while also building links with their destination). *Marginalization* appears when both destination attachment and home attachment are low, meaning that individuals lose connection to their home country and do not build new connections with their destination. The separation among the various types of acculturation can be done by defining strict thresholds on the two indices; however, we believe that they are more useful if considered as a continuum, as a means to place a person on a spectrum of acculturation rather than assigning a unique label.

The method cited here for studying acculturation has an important advantage in that it can be applied globally, in all countries where Twitter is used, and with space and time resolutions that can be arbitrarily low or high. This can complement analyses based on traditional data where resolution and coverage are limited by the complexity of the data-collection process. Additionally, various aspects of acculturation can be analysed separately. While all topics are discussed here, one could also concentrate only on political or other socially relevant topics to zoom in on very specific dimensions of acculturation.

The application of this method does not come without challenges or disadvantages. While Twitter data are public, downloading large amounts of data requires time, specific computer science skills and infrastructure. Data collection and analysis are further complicated by frequent changes to the Twitter API. Furthermore, the API offers only a subsample of the total data, which is further reduced by the fact that geolocation, required to assign residence and nationality of users, is present only in a minority of tweets. In addition, our final analysis is based only on users whose country of residence

is different from their country of nationality. As such, while initial Twitter data sets can be very large, multiple stages of filtration can result in rather small final data sets. Hence a lot of effort needs to be allocated to the data-collection phase, which should not be overlooked when planning an analysis.

Another challenge relates to privacy concerns. While using census or survey data to study migration is an established and regulated procedure, the Twitter platform is relatively new, and there are no established frameworks for using data from Twitter or other social media platforms for analysis of public policy issues. The European General Data Protection Regulation includes very strict rules for user profiling used to make automated decisions that target individuals and may have legal or other types of personal consequences. However, the analysis presented here does not have any component that involves individual decision-making. It is aimed at studying the process at the population level only, just as much as and in the same way integration indices are extracted from census data. Thus, researchers involved in any such analysis are not permitted to disclose any personal information, and all results are presented in a fully aggregated manner, in both space and time, similar to the publication of census data. Special caution should be exercised when studying smaller geographical areas: zooming in can create security concerns, so privacy evaluations should be undertaken.

Another important challenge with using social network data such as Twitter is selection bias. While Twitter is used globally, the distribution is not uniform. Some countries have broader coverage in the population compared to others, while in some countries, the platform is not used at all. Patterns of usage may also differ depending on individual socioeconomic characteristics. This means that results need to be interpreted carefully and not generalized before a proper analysis of the representativeness of the sample.

All in all, this case study shows that Twitter data can be successfully employed to study acculturation. Such analyses should be viewed as complementing and not alternate solutions to studies based on traditional data.

Measuring refugee integration through mobile phone data analysis

Mobile phones can provide detailed perspectives into the behaviour of people in a country through the mobility traces generated through their use (Ahas and Mark, 2005). Mobile phone data can also allow the gathering of insights about aspects of integration. Several ways exist to obtain useful information via mobile phones. An application installed on the phone by the user can access the GPS location (provided that the user has granted this access through the user agreement), which in turn can be used to identify home and work locations; track encounters, visited places and patterns of daily living; and know the socioeconomic status and many other indicators about the user. Given the highly sensitive character of these data, aggregation and desensitization – i.e. the removal of personal information – are essential to safeguard privacy and handle these data responsibly.

A possible source of information for measuring migrant and refugee integration are mobile phone applications specifically used by these population groups. However, the number of users and any biases in their composition must be taken into account when analysing such data. Mobile phones are like sensor packages carried by their users, and they periodically report the users' approximate locations

to the telecommunications company servicing them. This is necessary to provide the communication service in the first place. One consequence is that the mobile call detail records (CDRs, generated via calls and text messages) and extended detail records (XDRs, containing indications of data exchange) kept by telecommunication operators also contain traces of each user's movements. It is possible to use mobile CDRs for measuring refugee integration, provided that the CDR data are properly anonymized and aggregated (Salah et al., 2019a).

The most extensive initiative linking mobile CDRs to social integration has been the Data for Refugees (D4R) Challenge, which focused on Syrian refugees in Türkiye and opened an anonymized data set collected from one million users over an entire year to the research community (Salah et al., 2018; Salah et al., 2019a). The challenge prioritized five areas, of which social integration was the most popular topic among the participants.

What made D4R special among other data collaboratives using mobile CDRs was that records allowed to differentiate CDR data associated (with a high probability) with refugees from other users' data. It is important to note here that Türkiye is a party to the 1951 Refugee Convention, but the country recognizes "refugee" status only for people from Europe; the Syrian refugees were officially considered "temporarily protected foreign individuals". An indicator of possible refugee status was derived for customers in the database that (a) have ID numbers given to refugees and foreigners in Türkiye (42.87%), (b) are registered with Syrian passports (1.06%), and (c) use special tariffs reserved for refugees (56.07%). Roughly one fifth of the database referred to such individuals.

Several indicators based on CDR analysis can be interesting from an integration perspective. Collective mobility and behavioural similarity with locals can indicate integration for various reasons (Alfeo et al., 2019). These can show the extent of interactions between locals and migrants, illustrate economic capacity for migrants, as well as capture anomalies indicating social tension.

Various aspects of migrant integration can be analysed through specific indicators (Bakker et al., 2019). The *social integration* aspects measure social ties between the migrants (or refugees) and the host country. For each migrant user, a proxy for social integration can be the number of calls that are made to native-born users relative to the total number of calls made to all users. Since an indicator is needed in the data set to compute this proxy, it is relatively rare to see this information available for analysis. The second aspect is *spatial integration*, which derives from the usage of urban spaces by migrant and non-migrant populations. Mobile CDRs allow measuring proportions of these populations per area, as well as charting out of locations where native-born and migrant populations come together. For example, an index of dissimilarity can be computed for measuring spatial segregation, and an index of exposure can be calculated for measuring isolation (Iceland et al., 2002; Boy et al., 2019). Finally, *economic integration* relates to economic activities. To measure such activity, CDR data can help identify home/work locations (by the amount of activity within and outside working hours) and commuting patterns. The dissimilarity index, for example, can be computed separately for working hours and non-working hours to distinguish between residential segregation and employment segregation of the refugee and non-refugee populations in each province (Bertoli et al., 2019).

Several indicators can be directly computed via mobile data analysis. By modelling home attachment via mobile CDRs, one study found that in cities where the share of refugees in the total city population is high, there is a greater tendency to interact with citizens (Boy et al., 2019). Another way to study this aspect is to find locations where refugees and locals come together (Sterly et al., 2019). Compound indicators also exist, for instance based on CDR data merged with real estate market data and media data (Bertoli et al., 2019). Extensive experiments with mobility illustrate what data sources can be used as proxies. Home-based and non-home-based activities of refugees can be analysed with "point-of-interest data" (e.g. restaurants, schools, government offices) to determine integration levels per city (Hu et al., 2019). It is important to note that home and work locations – and similarly sensitive information – are not known for certain but calculated based on rough assumptions. Furthermore, individual-level analysis is never performed; only aggregated analysis is used, where data preprocessing ensures that no individuals can be identified or tracked using the mobile data. Ethical and privacy issues are discussed extensively in the article by Vinck et al. (2019).

Working with mobile CDRs involves significant challenges. Besides difficulties in accessing such data, CDRs contain data gaps. Children are mostly missing (as they are not legal owners of phone lines), and there are gender-related issues. For example, in D4R, a large proportion of the phone lines used by women are officially registered to a male (typically the husband or a son). CDRs indicate presence only during communication, thus individuals who rarely communicate are less represented. This is particularly important for migrants, who may have a smaller social network to contact compared to locals.

Discussion and policy implications

An ideal policy approach to migrant integration should reiterate informed decision-making and public support for comprehensive and reliable statistics. Driven by known gaps in quality and compatibility of migration data (Boy et al., 2019), alternative data sources and new methodologies for developing indicators for migratory movements and migrant integration offer great potential and are increasingly being explored.

Every time a mobile device or Internet service is used by individuals, data are being generated, stored and shared by private companies. This chapter illustrated how such data can be useful for migration policymakers and discussed both the potential and the limitations of using, in particular, mobile phone data and social media data for understanding different aspects of migrant and refugee integration. We demonstrated, using Twitter as a case study, that social media platforms are a promising source for measuring different dimensions of acculturation – namely, integration, separation, marginalization and assimilation. The main benefit of this approach is that it allows expanding the geographical coverage of the analysis, providing a wide application. Our second case study focused on the uses of mobile phone data for measuring refugee integration from both social and spatial perspectives. Extensive research in ethically responsible and privacy-compliant uses of mobile CDRs created new possibilities for initiation of private–public collaborations in this area (Verhulst and Young, 2019). Our case studies above illustrate that social media (Twitter) data and mobile phone data (CDRs) might be used to develop indicators for home attachment and social integration. Further, each case demonstrates other potential estimates and indicators such as spatial integration and resettlement (through CDRs), and different dimensions of acculturation (through Twitter data) can be developed using

big data analytics. Both these data sources, once the processing pipeline is established, can provide much timelier information than traditional data-collection approaches and expand the horizon of policymakers. While CDR data offer great coverage of mobility and migrant–non-migrant encounters across a country, Twitter provides opinion and sentiment analysis opportunities, providing insights about complementary aspects.

Following validity and ethical assessments, hands-on utilization of new data sources will serve the purpose of complementing existing (traditional) data sources with timelier and more accurate estimators. While such applications can complement the shortcomings of existing (traditional) data used to inform integration policies, such as underrepresented groups in migration data, interpretations are challenging since the concepts are more restricted when compared to steady and conventional methods. It is crucial to regulate how mobile phone and social media data will be processed and analysed. The existing schemes of aggregation and anonymization can ensure privacy protection and compliance with legal and ethical norms. Further collaboration between researchers, industry and policymakers on enhancing the use of artificial intelligence and machine learning for public policy decisions can prevent potential risks to fundamental rights, as described in a recent focus paper by the European Union Agency for Fundamental Rights (2018).

One of the challenges in such initiatives is that the technical concerns are often far from the policy issues, and a dialogue between data scientists and public authorities for specific cases is generally lacking (Salah et al., 2019b). However, new initiatives on ethically responsible data science, artificial intelligence for social good initiatives, technically capable subunits of intergovernmental organizations and non-governmental organizations, and more policy-aware governance of scientific projects are helping to bridge this gap.

Acknowledgements

Authors are supported by the European Commission through the following Horizon 2020 European projects: HumMingBird – Enhanced migration measures from a multidimensional perspective (GA: 870661), SoBigData Research Infrastructure – Social Mining and Big Data Ecosystem (GA: 654024), and SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics (GA: 871042).

REFERENCES*

Ahas, R. and Ü. Mark

2005 Location based services – New challenges for planning and public administration? *Futures*, 37(6):547–561.

Alfeo, A.L., M. GCA Cimino, B. Lepri, A.S. Pentland and G. Vaglini

2019 Assessing refugees' integration via spatio-temporal similarities of mobility and calling behaviors. *IEEE Transactions on Computational Social Systems*, 6(4):726–738.

Bakker, M.A., D.A. Piracha, P.J. Lu, K. Bejgo, M. Bahrami, Y. Leng, J. Balsa-Barreiro, J. Ricard, A.J. Morales, V.K. Singh, B. Bozkaya, S. Balcisoy and A. Pentland

2019 Measuring fine-grained multidimensional integration using mobile phone metadata: The case of Syrian refugees in Turkey. In: *Guide to Mobile Data Analytics in Refugee Scenarios* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 123–140.

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence and J. Weinstein

2018 Improving refugee integration through data-driven algorithmic assignment. *Science*, 359(6373):325–329.

Berry, J.W.

2019 *Acculturation: A Personal Journey across Cultures*. Cambridge University Press.

Bertoli, S., P. Cintia, F. Giannotti, E. Madinier, C. Ozden, M. Packard, D. Pedreschi, H. Rapoport, A. Sîrbu and B. Speciale

2019 Integration of Syrian refugees: Insights from D4R, media events and housing market data. In: *Guide to Mobile Data Analytics in Refugee Scenarios* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 179–199.

Boy, J., D. Pastor-Escuredo, D. Macguire, R. Moreno Jimenez and M. Luengo-Oroz

2019 Towards an understanding of refugee segregation, isolation, homophily and ultimately integration in Turkey using call detail records. In: *Guide to Mobile Data Analytics in Refugee Scenarios* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 141–164.

European Union Agency for Fundamental Rights (FRA)

2018 #BigData: Discrimination in Data-supported Decision Making. Vienna.

** All hyperlinks were working at the time of writing this report.*

Hu, W., R. He, J. Cao, L. Zhang, H. Uzunalioglu, A. Akyamac and C. Phadke
2019 Quantified understanding of Syrian refugee integration in Turkey. In: *Guide to Mobile Data Analytics in Refugee Scenarios* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 201–221.

Iceland, J., D.H. Weinberg and E. Steinmetz
2002 *Racial and Ethnic Residential Segregation in the United States: 1980–2000*. Volume 8. United States Government Printing Office, Washington, D.C.

Kim, J., A. Sîrbu, F. Giannotti and L. Gabrielli
2020 Digital footprints of international migration on Twitter. In: *International Symposium on Intelligent Data Analysis*. Springer, pp. 274–286.

Kim, J., A. Sîrbu, G. Rossetti, F. Giannotti and H. Rapoport
2021 Home and destination attachment: Study of cultural integration on Twitter. arXiv preprint arXiv:2102.11398.

Magdy, A., T.M. Ghanem, M. Musleh and M.F. Mokbel
2014 Exploiting geo-tagged tweets to understand localized language diversity. In: *Proceedings of Workshop on Managing and Mining Enriched Geo-spatial Data*, pp. 1–6.

Monroe, T.
2018 *Big Data Solutions in Forced Migration: Innovations in Analytics to Promote Humane, Sustainable Responses to Forced Migration*. World Bank Group, Washington, D.C.

Salah, A.A., A. Pentland, B. Lepri and E. Letouzé (eds.)
2019a *Guide to Mobile Data Analytics in Refugee Scenarios*. Springer Nature Switzerland AG, Cham.

Salah, A.A., M. Tarık Altuncu, S. Balcisoy, E. Frydenlund, M. Mamei, M. Ali Akyol, K. Yavuz Arslanlı, I. Bensason, C. Boshuijzen-van Burken, P. Bosetti, J. Boy, T. Bozcaga, S. Mümin Cilasun, O. Işık, S. Kalaycıoğlu, A. Seyyide Kaptaner, I. Kayi, Ö. Ozan Kılıç, B. Kjamili, H. Kucukali, A. Martin, M. Lippi, F. Pancotto, D. Rhoads, N. Sevencan, E. Sezgin, A. Solé-Ribalta, H. Sterly, E. Surer, T. Taşkaya Temizel, S. Tümen and I. Uluturk
2019b Policy implications of the D4R challenge. In: *Guide to Mobile Data Analytics in Refugee Scenarios* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 477–495.

Salah, A.A., A. Pentland, B. Lepri, E. Letouzé, P. Vinck, Y.A. de Montjoye, X. Dong and O. Dagdelen
2018 Data for refugees: The D4R challenge on mobility of Syrian refugees in Turkey. arXiv preprint arXiv:1807.00523.

Sîrbu, A., G. Andrienko, N. Andrienko, C. Boldrini, M. Conti, F. Giannotti, R. Guidotti, S. Bertoli, J. Kim, C.I. Muntean, L. Pappalardo, A. Passarella, D. Pedreschi, L. Pollacci, F. Pratesi and R. Sharma
2021 Human migration: The big data perspective. *International Journal of Data Science and Analytics*, 11:341–360.

Sterly, H., B. Etzold, L. Wirkus, P. Sakdapolrak, J. Schewe, C.F. Schleussner and B. Hennig
2019 Assessing refugees' onward mobility with mobile phone data – A case study of (Syrian) refugees in Turkey. In: *Guide to Mobile Data Analytics in Refugee Scenarios* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 251–263.

Verhulst, S.G. and A. Young
2019 The potential and practice of data collaboratives for migration. In: *Guide to Mobile Data Analytics in Refugee Scenarios* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 465–476.

Vinck, P., P.N. Pham and A.A. Salah
2019 "Do no harm" in the age of big data: Data, ethics, and the refugees. In: *Guide to Mobile Data Analytics in Refugee Scenarios* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 87–99.

ADDITIONAL READING

Bircan, T., D. Purkayastha, A.W. Ahmad-Yar, K. Lotter, C. Dello Iakono, D. Göler, M. Stanek, S. Yilmaz, G. Solano and Ö. Ünver
2020 Gaps in migration research: Review of migration theories and the quality and compatibility of migration data on the national and international level. HumMingBird, Leuven.

Calderón, C.A., F. Ortega Mohedano, M. Álvarez and M. Vicente Mariño
2019 Distributed supervised sentiment analysis of tweets: Integrating machine learning and streaming analytics for big data challenges in communication and audience research. *Empiria: Revista de Metodología de Ciencias Sociales*, 42:113–136.

Franklinos, L., R. Parrish, R. Burns, A. Caflisch, B. Mallick, T. Rahman, V. Routsis, A.S. López, A. Tatem and R. Trigwell
2020 Key opportunities and challenges for the use of big data in migration research and policy. *UCL Open: Environment Preprint*.

Wimark, T., K. Haandrikman and M.M. Nielsen
2019 Migrant labour market integration: The association between initial settlement and subsequent employment and income among migrants. *Geografiska Annaler: Series B, Human Geography*, 101(2):118–137.

8

SOCIAL MEDIA

ARTIFICIAL INTELLIGENCE/ MACHINE LEARNING

MONITORING PUBLIC SENTIMENT TOWARDS MIGRANTS

USING TWITTER DATA TO MONITOR IMMIGRATION SENTIMENT

Francisco Rowe,[1] *Michael Mahony,*[2] *Eduardo Graells-Garrido,*[3] *Marzia Rango*[4] *and Niklas Sievers*[5]

Making a case for Twitter data

Monitoring sentiment towards migrants is a key ingredient to migrant integration and social cohesion. Immigration has consistently been identified as one of the most divisive social issues globally (EPSC, 2019). Immigration sentiment shapes migration policy formulation and political outcomes. Anti-immigration sentiment has spurred attention towards more restrictive migration policies, and it has been linked to an increasing prominence of right-wing affiliation, particularly in Western European countries and the United States of America (Bail et al., 2018; Greven, 2016; Dennison and Geddes, 2019). Immigration sentiment also influences the capacity of migrants to successfully participate in activities in their receiving communities. Acts of discrimination, intolerance and xenophobia can impair immigrants' ability to secure employment and housing, and achieve a sense of belonging in local communities, contributing to more polarized societies (Blinder and Richards, 2020; Cheong et al., 2007; Penninx et al., 2008).

Anti-immigration sentiment has gained salience during the last decade. A wave of key politically motivated events has significantly contributed towards this trend. Examples include the Brexit referendum in the United Kingdom and Donald Trump's presidential campaign in the United States. More recently, the world has faced an unprecedented challenge to tackle and understand the spread and impacts of COVID-19, which

[1] Francisco Rowe is Lead of the Geographic Data Science Lab and Senior Lecturer in Quantitative Human Geography of the Department of Geography and Planning at the University of Liverpool. He holds a PhD in Economic Geography from the University of Queensland in Brisbane, Australia, and an MSc in Regional Science from the Universidad Católica del Norte in Antofagasta, Chile. Francisco is the corresponding author and can be contacted at F.Rowe-Gonzalez@liverpool.ac.uk.

[2] Michael Mahony is a PhD student of the Geographic Data Science Lab at the University of Liverpool. His work seeks to identify representative labour market trajectories of immigrants, and key individual, household and contextual factors underpinning these trajectories. He has a bachelor's degree in Land Economy from the University of Cambridge and an MSc in Geographic Data Science from the University of Liverpool.

[3] Eduardo Graells-Garrido is Professor and Researcher at the Instituto Data Science of the Universidad del Desarrollo, Chile. Previously he worked as a researcher on mobility at the Barcelona Supercomputing Center, Spain. He holds a PhD in Information and Communication Technologies at the Universitat Pompeu Fabra, Spain.

[4] At the time of writing, Marzia Rango was leading the work on data innovation, capacity-building and analytics at IOM's Global Migration Data Analysis Centre (GMDAC) in Berlin. She is the co-convenor of the Big Data for Migration Alliance, a joint initiative of GMDAC, the European Commission's Joint Research Centre and the Governance Lab at New York University, seeking to accelerate the responsible use of new data sources and methods for migration analysis and policy. She now works as a Migration and Human Mobility Specialist at the United Nations Operations and Crisis Centre in New York.

[5] At the time of writing, Niklas Sievers is Data Knowledge Officer at IOM GMDAC. He developed several research projects exploring the potential and pitfalls of innovative data sources, methods, and tools for migration policy and research. Before joining the United Nations System, he worked as an Advisor in the European Union Advisory Unit of PricewaterhouseCoopers and as a Lecturer at Humboldt University. He holds an MSc from the London School of Economics and Political Science in the social sciences and two bachelor's degrees from Leuphana University.

reportedly coincided with an increase in anti-immigration sentiment (*Nature*, 2020). Acts and displays of intolerance, discrimination, racism, xenophobia and violent extremism emerged, linking individuals of Asian descent and appearance to COVID-19 (ibid.).

Understanding public opinion towards immigrants is key to preventing the spread of misinformation that fuels misperception, negative attitudes and discrimination against immigrants. Anti-immigration sentiment is often rooted in misperceptions (EPSC, 2019), and experimental evidence has revealed that providing information to address these misconceptions can shift attitudes towards a more supportive view of immigration (Grigorieff et al., 2020). The availability and accuracy of data on public opinion about migration are thus critical for tackling misperceptions and understanding the extent of local openness to immigration and ethnic diversity (Dennison and Dražanová, 2018).

Traditionally, data on public attitudes towards immigration are collected through qualitative sources, namely ethnographies, interviews and surveys. Yet qualitative methods rely on small samples and normally suffer from sample bias (Atieno, 2009). Similarly, while surveys can provide a reliable national representation, they are expensive and infrequent, offer low population coverage, lack statistical validity at fine geographical scales, and become available with a lag of one or two years after they have been collected (ibid.). Additionally, survey data do not normally provide insights into why people hold certain views on migration, and respondents may interpret the same survey question differently (Goyder, 1986).

New forms of data provide an opportunity to overcome these deficiencies. Social media, particularly microblogging, offers a dynamic and open space which provides a unique window to better understand public opinion about immigration. Microblogging is a new form of communication in which users can publish short posts to express live opinions on mobile phones, computers and tablets. In January 2021, 54 per cent (4.2 billion) of the world's population were estimated to be active social media users (Hootsuite and We Are Social, 2019). Social media data are produced at an unprecedented temporal frequency and geographical granularity, and they are accessible in real time (McCormick et al., 2017). Coupled with cheap computing and machine-learning algorithms, these data enable real-time processing of information to measure and monitor anti-immigration sentiment at frequent temporal intervals over extended time frames and across the globe (Bartlett and Norrie, 2015; Freire and Graells-Garrido, 2019).

This paper aims to illustrate how immigration sentiment can be measured and monitored using Twitter data and natural language processing (NLP). Drawing on Twitter data, we examine immigration sentiment in the United Kingdom from 15 January to 15 February 2020, comprising the start of the unfolding COVID-19 pandemic. The first reported case of COVID-19 in the United Kingdom dates back to 31 January 2020 (Wikipedia, n.d.a), a day after the first reported media case of a COVID-19-related incident of racism and xenophobia against a Chinese student on 30 January 2020 (Wikipedia, n.d.b).

The potential limitations of Twitter in capturing immigration sentiment should be acknowledged. This source may capture the opinions only of a selected segment of the population whose size and attributes vary by country, according to access to digital technology, offering a partial representation

of immigration sentiment. Yet Twitter represents a novel, promising source of data that could complement traditional sources or offer valuable data insights where appropriate data are not readily available. Thus, Twitter data can contribute to developing a more timely and comprehensive understanding of public perceptions towards immigration.

Complementing survey research with Twitter data

Globally, surveys are the primary source of public opinion on migration. A number of surveys provide historical data on attitudes towards immigration, including the Gallup World Poll, the Pew Global Attitudes Project, the International Social Survey Programme, the World Values Survey, the Ipsos Global Trends, the European Social Survey and the Eurobarometer. They vary widely in temporal frequency, the number of countries covered and the number of questions collecting data on migration opinion (IOM, 2017). A Gallup World Poll to collect data on attitudes towards immigration was conducted between 2012 and 2014 and covered a total of 160 countries; however, it contained only six questions on migration-related public opinion, and data are not publicly accessible (ibid.). The World Values Survey has run annually since 1981 but typically contains only two questions on migration-related public opinion, while the European Social Survey contains 36 of such questions, but these were used only in two waves in 2002 and 2014 (ibid.).

A common question to capture broad attitudes towards immigration is whether or not people think immigration levels should increase, decrease or stay at present levels. While wide variability across countries exists, world regional averages based on Gallup World Poll data collected in 2012–2014 revealed people's preference for either maintaining or increasing current immigration levels, with Europe being an exception (Esipova et al., 2015). In Europe, residents appeared to have the least positive attitudes towards immigration globally, with 52 per cent of the surveyed population indicating a need to reduce current immigration levels (ibid.). Yet a sharp divergence emerged between Northern and Southern Europe (ibid.). Southern Europeans tended to display more negative attitudes towards immigration, preferring lower national immigration levels, while Northern Europeans showed more positive attitudes, favouring maintenance or an increase in current immigration levels (Dennison and Geddes, 2019; Esipova et al., 2015). In Northern Europe, the United Kingdom stood out as an exception with a larger percentage of the population in favour of lower immigration levels (Esipova et al., 2015).

Nonetheless, anti-immigration sentiment in Europe seems to be softening. Longitudinal data on feelings towards immigrants from Eurobarometer surveys conducted in 2014–2018 reveal a decreasing trend in anti-immigration sentiment across most European countries (Dennison and Geddes, 2019). Though, stronger negative feelings exist towards immigrants from non-European Union nations than from European Union member States (ibid.). In the United Kingdom, a paradoxical softening in anti-immigration sentiment has taken place since the Brexit referendum (Schwartz et al., 2021). According to Ipsos data, the share of the population agreeing that there are too many migrants has reduced from 64 per cent in 2013 to 45 per cent in 2017 (Blinder and Richards, 2020). Additionally, there is now a predominantly positive perception of the impact that immigration has had on Britain (Ipsos, 2019).

While existing survey data on public opinion about immigration provide a valuable understanding of long-term changes in attitudes towards immigration and cross-national differences, challenges remain. Survey data are typically spatially coarse, costly and infrequent. Existing data can be limited through slow data releases and statistical representation, especially at small geographical units. Real-time, frequent, exhaustive and internationally spanning information is crucial to monitor changing attitudes towards immigrants during dynamic and fast-evolving events, such as pandemics. Twitter data offer a novel source to complement traditional data systems and cover their gaps and feed into near real-time monitoring of immigration sentiment.

The potential of Twitter data is reflected in three key areas: (a) high geographical granularity, (b) real-time temporal frequency and (c) global coverage. Traditional survey data are typically available at the national or administrative level. Twitter data are individually geolocated and can thus be aggregated at convenient geographic units of analysis best representing the social process of concern. Twitter data also offer a high degree of temporal frequency in real time. Twitter data are timestamped, recording information across the temporal continuum, comprising years, months, days, minutes and seconds. Such rich temporal granularity in real time enables the tracking of online discussions (Wang et al., 2012), diseases (Paul and Dredze, 2012) and natural disasters (Bruns and Liang, 2012), as well as the management of emergency responses (Terpstra et al., 2012). Moreover, Twitter provides a worldwide repository to analyse the global patterns of human mobility (Hawelka et al., 2014), misinformation (Vosoughi et al., 2018) and emotion (Larsen et al., 2015). However, these promises are impacted by key caveats. Only a fraction of tweets contain spatial information. Less than 3 per cent of tweets are geolocated (Twitter, n.d.). The global representation of Twitter data is affected by the demographic profile of Twitter users (Leetaru et al., 2013), along with the censorship of Twitter in China, which is the world's most populous country. Nevertheless, Twitter data represent a novel and valuable resource to complement traditional survey data.

Describing the data and methods

Collecting and engineering Twitter data

A first key task is to collect data from Twitter. Two key elements form the data-collection strategy: (a) method of data collection and (b) a clearly defined set of search terms. A first critical step that should be considered to define an effective data-collection strategy is the method of data collection. Twitter data are collected through an application programming interface (API). As described in Section 5 ("Key consideration for practitioners"), Twitter provides two different APIs, each offering different capabilities in terms of data allowance and access. The selection of an API will thus determine the type of data that can be accessed. The second important step is to define a clear set of search terms, to guide the content of the tweets to be retrieved. Carefully considering the various terms that may be used to discuss a given topic is key to developing a comprehensive search and data-collection strategy.

We draw on a random sample of 1.76 million tweets from the United Kingdom covering the start of the COVID-19 pandemic, between 15 January and 15 February 2020. This sample comprises a sample of 22,000 original tweets and 1.73 million retweets, reflecting the prevalence of retweets

about this topic. Data were collected via an API (Campan et al., 2018). We used Twitter's premium API to access historical data with a monthly cap of 1.25 million tweets. It enables 500 tweets per request at a rate of 60 request per minute; access to tweets, retweets, URLs, hashtags and profile geographic information; and a total number of 2,500 requests per month.

The data were collected based on a random sampling strategy. To maximize our monthly API data allowance, a sampling strategy was developed to collect a sample of 1,500 tweets on a daily basis from 1 December 2019 to 31 June 2020. We generated a data set for a larger project to monitor immigration sentiment during the course of the COVID-19 pandemic across five countries, including the United Kingdom (Rowe et al., 2021). We collected data at the peak hour of daily tweet activity using a geographic bounding box. We assessed the statistical representation of the resulting data set, comparing sentiment scores based on four lexicons against scores obtained from a data set containing all daily tweets for seven full days. The resulting sentiment scores from both data sets were consistent, identifying similar daily patterns of immigration sentiment.

To collect tweets focusing on migration, we were guided by the principles of the Campbell policies and guidelines for the conduct of systematic reviews (Campbell Collaboration, n.d.). A key component of conducting a systematic review is planning a search strategy to capture relevant content. In consultation with migration experts at IOM, we developed a set of key search terms, including words, Twitter accounts and hashtags. Table 1 lists the selection of words and hashtags included in our search terms. Twitter accounts are not displayed for privacy and confidentiality purposes.

Table 1. Search terms used for tweet data collection

A key component of conducting a systematic review is planning a search strategy to capture relevant content. In consultation with migration experts at IOM, a set of key search terms were developed, including words, Twitter accounts and hashtags.

Categories		Search Terms
Terms		immigrant, immigration, migrant, migration, "asylum seeker", refugee, "undocumented worker", "guest worker", "EU worker", "non-UK workers", "foreign worker", (human smuggling), (human trafficking), illegals, foreigner, "illegal alien", "illegal worker", islamophob*, sinophob*, "china flu", "kung flu", "china virus", "chinese virus", shangainese
Accounts		@UNmigration, @IOM_UN, @IOMatUN, @IOMatEU, @IOM_UK, @IOMResearch, @IOM_GMDAC, @hrw, @Right_to_Remain, @CommonsHomeAffs, @fcukba, @Mark_George_QC, @MigrantVoiceUK, @MigrantChildren, @MigrantHelp, @thevoiceofdws, @WORCrights, @UbuntuGlasgow, @MigrantsUnionUK, @migrants_rights, @MigrantsMRC, @Consenant_UK, @RomaSupport, @MigrantsLawProj, @MigRightsScot, @IRMOLondon, @HighlySkilledUK, @WeBelong19, @Project17UK
Hashtags	Positive	@ukhomeoffice, @pritipatel, @UKHomeSecretary, @EUHomeAffairs, @MigrMatters, @MigObs
	Neutral	@UNmigration, @IOM_UN, @IOMatUN, @IOMatEU, @IOM_UK, @IOMResearch, @IOM_GMDAC, @hrw, @Right_to_Remain, @CommonsHomeAffs, @fcukba, @Mark_George_QC, @MigrantVoiceUK, @MigrantChildren, @MigrantHelp, @thevoiceofdws, @WORCrights, @UbuntuGlasgow, @MigrantsUnionUK, @migrants_rights, @MigrantsMRC, @Consenant_UK, @RomaSupport, @MigrantsLawProj, @MigRightsScot, @IRMOLondon, @HighlySkilledUK, @WeBelong19, @Project17UK, @ukhomeoffice, @pritipatel, @UKHomeSecretary, @EUHomeAffairs, @MigrMatters, @MigObs, @Nigel_Farage, @MigrationWatch
	Negative	#illegals, #foreigner, #foreigners, #illegalalien, #illegalaliens, #illegalworker, #OurCountry, #illegalworkers, #KeepThemOut, #SendThemBack, #migrantsnotwelcome, #refugeesnotwelcome, #illegals, #ChinaVirus, #chinaflu, #kungflu, #chinesevirus, #TheyHaveToGoBack, #DeportThemAll
	Event	#Moria, #CampFire, #closethecamps

Using sentiment analysis

To capture immigration sentiment, we used sentiment analysis, also known as opinion mining or emotion artificial intelligence. This refers to the use of NLP to systematically identify, measure, and analyse emotional states and subjective information. It computationally enables the polarity of text to be identified – that is, whether the underpinning semantics of an opinion is positive, negative or neutral. Furthermore, it allows deriving quantitative scores to identify the attitude or position of a given piece of text based on the distribution of negative or positive terms in it.

We specifically employed VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert, 2014). VADER is a lexicon and rule-based sentiment analysis tool which is tailored to the analysis of sentiments expressed in social media. VADER has been shown to perform better than 11 typical state-of-the-practice sentiment algorithms at identifying the polarity expressed in tweets (ibid.). It overcomes limitations of existing approaches by more appropriately handling informal text, including the use of negations, contractions, slang, emoticons, emojis, initialisms, acronyms, punctuation, and word shape (e.g. capitalization) as a signal of sentiment polarity and intensity. Most commonly, lexicon-based approaches capture differences only in sentiment polarity (i.e. positive or negative); they do not identify differences in sentiment intensity (strongly positive versus moderately positive) or contradictions (e.g. "Immigration is good, but the current visa system is *horrible*"). They have also been designed to capture sentiment in well-structured sentences – meaning, generally their lexicons do not include slang, emoticons, emojis, acronyms and capitalized word differentiation. We note that accurate identification and scoring of sarcastic statements remain a key challenge in NLP, but these statements tend to represent a small fraction of daily tweets.

VADER provides a normalized, weighted composite score which captures the polarity and intensity of individual tweets. The score ranges from -1 to +1, representing the most extreme negative to the most extreme positive sentiment respectively. Intuitively, to derive the score, VADER assigns a value to each word in a tweet, ranging from -4 (extremely negative) through 0 (neutral) to +4 (extremely positive) based on positive and negative text features identified in the post. These scores are then aggregated and normalized to range between -1 and +1. We used the daily average of the composite score to track the daily evolution of immigration sentiment on Twitter. We then identified positive sentiment tweets (i.e. composite score > 0.05) and negative sentiment tweets (i.e. composite score < -0.05).

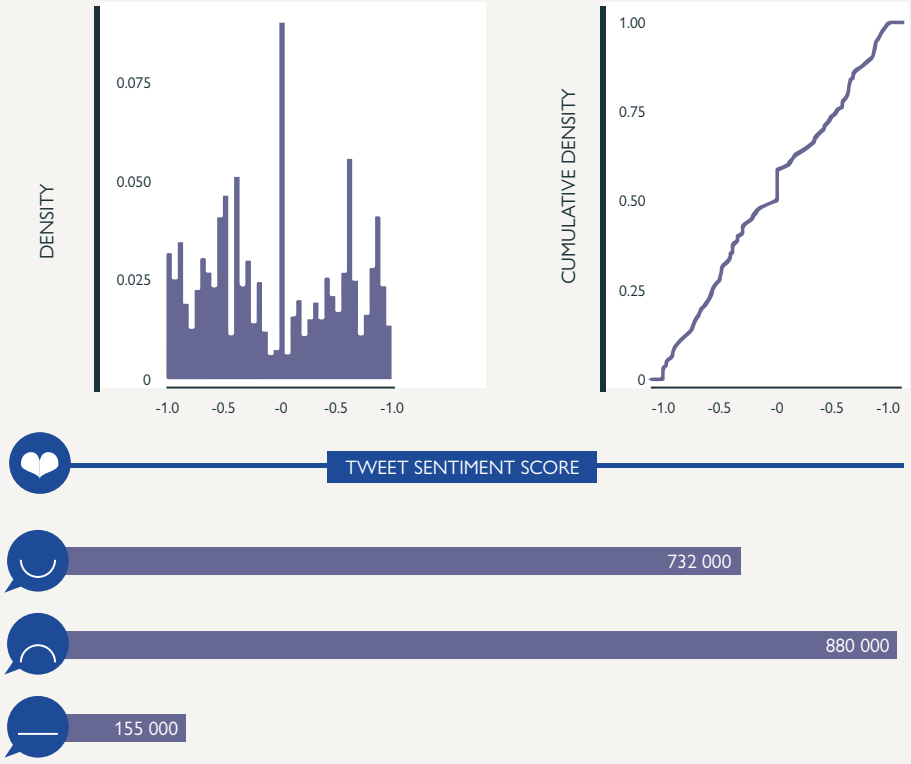
Gaining valuable insights from Twitter data

Understanding the distribution of sentiment

Figure 1 displays the overall distribution of tweet sentiment scores between 15 January and 15 February 2020. As expected, it shows a high frequency of neutral polarity tweets but also reveals a high prevalence of negative and positive polarity scores around -0.5/-1 or 0.5/1, indicating the existence of a very polarized discussion on issues relating to immigration. Adding together all negative sentiment scores results in a total of 880,000, which exceeds the number of positive sentiment tweets (723,000), while neutral sentiment tweets account for a small fraction (155,000). These results reflect the fact that Britain has become an increasingly divided society on controversial issues, and immigration has featured as a key divisive topic (Blinder and Richards, 2020), particularly prominent

during the lead-up to the Brexit referendum (Bruns and Liang, 2012). These divisions have become increasingly aligned with partisan identities in recent years (Schwartz et al., 2021) and have been attributed to echo chambers – patterns of information-sharing that reinforce pre-existing beliefs by restricting exposure to opposing political views (Bail et al., 2018). Social media is often believed to comprise a main channel leading to selective exposure to information and political polarization (Conover et al., 2011; Hong and Kim, 2016).

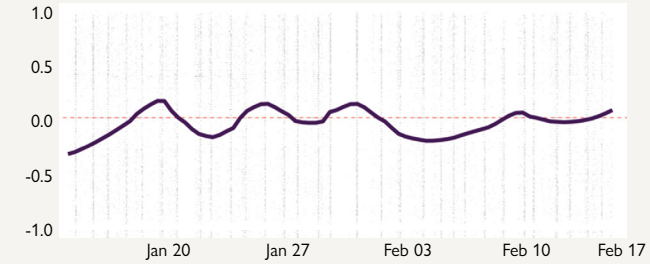
Figure 1. Density (left) and cumulative (right) distribution of sentiment scores



Examining temporal variations

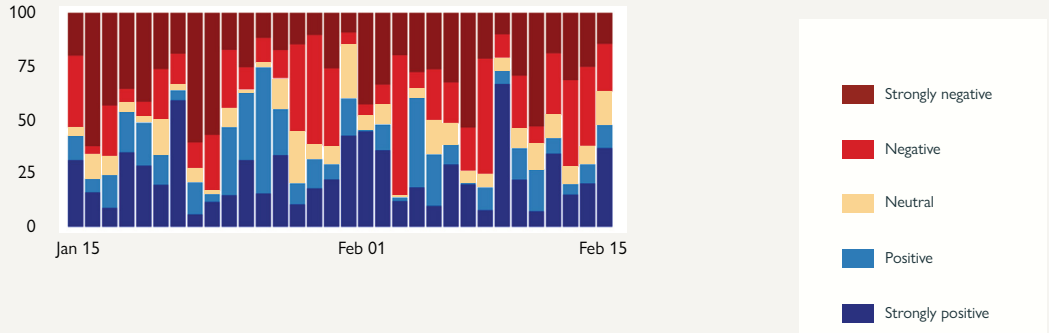
Figure 2a shows the daily average tweet sentiment score displaying a cyclical pattern of positive and negative sentiments, and Figure 2b reveals the sentiment intensity composition of tweets. A first key observation from these results is that negative feelings towards immigration do not seem to have intensified during the start of the COVID-19 pandemic in the United Kingdom. This is contrary to expectations of increased intensification of negative sentiments towards immigration, particularly against people of Asian descent and appearance. Links between COVID-19 and China are believed to have sparked acts and displays of xenophobia and racism against Asian people around the world (*Nature*, 2020). Asian international students were reportedly subjected to racist attacks in the early stages of the outbreak (ibid.).

Figure 2. Daily evolution of tweet sentiment



Average overall sentiment score. Smoothed conditional means obtained are reported obtained via locally weighted scatterplot smoothing (LOESS) using a span of 0.3. Gray background dots represent daily individual tweet sentiment scores.

Percentage of sentiment scores classified into strongly negative (< -0.5), negative (-0.5 to -0.05), neutral (-0.05 to 0.05), positive (0.05 to 0.5) and strongly positive (> 0.5).



Unpacking daily discussions

Figure 2a suggests that key Brexit events have driven periods of intensification in negative sentiment. Notable periods of negative sentiment include 15 and 16 January (initial days in our analysis), 22 and 23 January, and 3 to 8 February 2020. On 15 and 16 January, these dates coincide with concerns raised by the European Parliament on 15 January about European Union citizens risking discrimination after Brexit in seeking housing and employment (Rankin, 2020). On 22 and 23 January, these dates concur with the Government's unveiling of its post-Brexit immigration plans (*The Guardian*, 2020). The dates from 3 to 8 February comprise the first working week after the United Kingdom exited the European Union on Friday, 31 January; Trump's impeachment acquittal on 5 February; and the indictment of multiple murder charges against an immigrant in the United States on 7 February (Smith, 2020; Li, 2019). Figure 2b reveals that overall negative sentiments during these three sets of dates were driven by a rise in the percentage of tweets with strongly negative sentiment, with relatively less tweets expressing strongly positive sentiment. This situation contrasts with days during which relatively high percentages of negative sentiment tweets do not seem to result in an overall negative sentiment day score, as these tweets are counteracted by an equally large share of strongly positive sentiment tweets, such as on 13 February.

To better understand periods of rising negative sentiment, we analyse the frequency of words during the three sets of dates identified above. We distinguished between words associated with expressions of negative and positive feelings captured in tweets, to identify general themes of discussion. A main advantage of Twitter data is its temporal frequency, which enables understanding of changes in public opinion on a daily basis and the ways they are shaped by key events. Figure 3 displays the top 25 most frequently used words in tweets associated with negative (red) and positive (blue) sentiments on 15 and 16 January, 22 and 23 January, and from 3 to 8 February 2020. Words relating to COVID-19

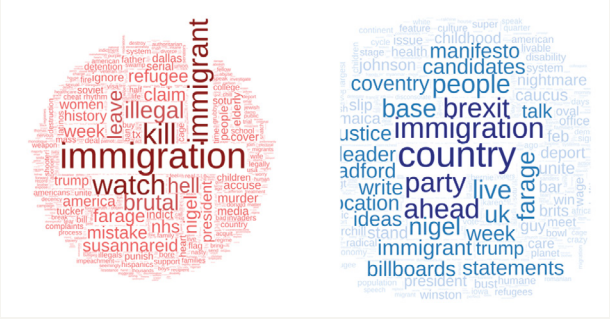
are not apparent. Crime and the implications of Brexit and the resulting immigration policy on European Union migration, refugee settlement, and human trafficking emerge as key themes from the analysis of words frequency.

Figure 3. Frequent words to express negative (red) or positive (blue) sentiment

KEY THEMES EMERGED:
Crime, the implications of Brexit and immigration policy on European Union migration, refugee settlement and human trafficking.

(a) On 15–16 January 2020

(b) On 22–23 January 2020



(c) On 3–8 February 2020

Negative sentiment

Positive sentiment

Words such as “alien”, “deport”, “kill” and “arrest”, related to a criminal event, represent the prevailing negative sentiment on 15 and 16 January. They relate to a viral tweet shared by more than 13,900 users involving a notorious case of assault and murder by an apparently undocumented immigrant in the city of New York in the United States. The consequences of Brexit for European Union citizens and Britons are a prevalent topic across the three sets of dates. Negative feelings are addressed to the hard realization of the challenges imposed by Brexit. Tweets express the anger and frustration of European Union citizens and Britons as they are expected to apply for residency to remain in their respective host countries, where they pay taxes and own a house. Tweets relating to Brexit also embody positive sentiment as captured in the word clouds, celebrating the exit of the United Kingdom from the European Union and how this enables the controlling of immigration. Stringent

policy on human smuggling and refugee issues are prevalent themes as well, and they seem to be key triggers of both positive and negative sentiments. On 22 and 23 January, tweets revolve around the rejection of amendments to the child refugee policy by the United Kingdom's Government. From 3 to 8 February, tweets relate to temperamental comments from people coming to the realization of false claims made by the Leave campaign about 350 million pounds a week becoming available for the National Health Service post-Brexit, Trump's impeachment acquittal, and the murder indictment of an immigrant in the United States.

Key consideration for practitioners

Understanding public attitudes towards immigration is essential in the implementation of the Global Compact for Safe, Orderly and Regular Migration, specifically its Objective 17: "Eliminate all forms of discrimination and promote evidence-based public discourse to shape perceptions of migration." In the early stages of the COVID-19 outbreak, concerns about a rising number of racist and xenophobic incidents against individuals of Asian descent and appearance were reported. However, lack of appropriate data has prevented a detailed analysis of immigration sentiment since the start of the pandemic. This paper illustrates how Twitter data can be employed to analyse immigration sentiment during the start of the COVID-19 pandemic. We analysed a sample of 1.76 million tweets from the United Kingdom between 15 January and 15 February 2020. While Twitter data can be a powerful resource to measure and monitor changes in public opinion towards migrants, practitioners should carefully address the following key considerations.

Search strategy

Collecting Twitter data on immigration is challenging. A clearly defined search strategy is required. A wide variety of terms are used to describe issues related to migration on Twitter. Discussions may revolve around migration as a process or migrants themselves, and different expressions may be used. Conventional expressions such as "immigrants" and "migration" are often used to describe these discussions, but less conventional expressions like "Shanghainese" are also widely employed. A carefully curated collection of search terms is needed to comprehensively cover discussion relating to migration on social media. Rowe and team (2021) provide a curated list of terms to collect social media data on migration, comprising words, hashtags and accounts in four different languages.

Statistical representation

Ensuring representation of the populations under study is key to enable statistical inference. Yet we know particular groups of the population are overrepresented in Twitter data, and significant variability may exist across geographic areas within and between countries (Leetaru et al., 2013). Additionally, as indicated in Section 2 ("Complementing survey research with Twitter data"), the spatial representation of Twitter data is limited as less than 3 per cent of tweets are geolocated (Twitter, n.d.). Addressing these biases is an active area of research. Specific-case weighting schemes have been developed to ensure statistical and spatial representation of social media data (Grow et al., 2021).

Data access

While Twitter represents a global repository of virtual social interactions, constraints to data accessibility exist. Twitter allows data collection via APIs. Twitter has traditionally offered two APIs: streaming API and search API (Campan et al., 2018). The streaming API enables data collection in

real time, while the search API is used to retrieve historical data. The search API includes three tiers: streaming, premium and enterprise. Standard streaming is free and allows collecting a random sample of recently published public tweets in the past seven days, with coverage varying between 1 and 40 per cent of all tweets over time (Morstatter et al., 2013). The premium and enterprise options are paid and expensive, and they have a monthly tweet and per-request cap. A careful evaluation of these data accessibility constraints is thus required to develop an effective data-collection strategy. On 26 January 2021, Twitter introduced the academic research API, which is expected to greatly facilitate access to data. The academic research API is free, provides access to the full history of public conversation, and has a higher monthly tweet volume cap of 10 million (Tornes, 2021).

Building sentiment scores

Significant progress has been made on sentiment analysis, but key challenges remain. As described in Section 3.2 ("Using sentiment analysis"), VADER has contributed to overcoming many challenges by the use of negations, contractions, slang, emoticons, emojis, initialisms, acronyms, punctuation, and word shape and returning a score that captures the intensity of the sentiment. It is not a binary score capturing positive or negative sentiment. Yet challenges remain. A key challenge relates to the meaning of sentiment scores. Negative (or positive) sentiment scores do not necessarily represent a negative (or positive) comment against (or in favour of) migration. A comment linked to a negative sentiment score may contain negative expressions about visas or migration processes but a positive opinion about migrants or migration impacts. Our results, for instance, link migration discussions revolving around Brexit to both positive and negative sentiment scores. A second key challenge is that most sentiment analysis algorithms are trained on English text. Though some algorithms offer non-English dictionaries, translation is often required. Manual translation is expensive, and automated translation is not 100 per cent accurate. However, lexicon-based sentiment analysis algorithms like VADER perform well on automated translated text as they are trained and assign sentiment scores on individual words (Hutto and Gilbert, 2014). Algorithms trained on structured text, such as the Stanford NLP, work less accurately according to our experimentation. A third key challenge is handling sarcasm in tweets. Lexicon-based algorithms have difficulties in recognizing sarcastic comments. As indicated above, current algorithms are trained on individual words. Greater context may be needed to improve the efficacy of algorithms in identifying sarcasm. This may be possible by linking tweet replies and quotes to original tweets, but this cannot occur for original tweets. Though, according to our experience, sarcasm normally accounts for a small percentage of tweets and relates to image-based memes, not text.

Ethics

Twitter data offer new opportunities but also represent key ethical challenges. Twitter data can be described as identifiable. Regulatory frameworks and ethical guidance are thus required to access, manage and analyse data responsibly. Traditional institutional boards may not be suitable given the high level of technical knowledge required to understand the potential for intrusion and individual harm of new forms of data, and come up with appropriate ways of data anonymization. A principles-based approach has been proposed for the use of digital data in the social sciences (Salganik, 2019) – that is, the design and application of project-specific rules in more general context of ethical principles. A principles-based approach should help researchers make appropriate decisions for cases where rules have not yet been written, and guide the way in which research is communicated.

REFERENCES*

Atieno, O.P.
2009 An analysis of the strengths and limitation of qualitative and quantitative research paradigms. *Problems of Education in the 21st Century*, 13(1):13–38.

Bail, C.A., L.P. Argyle, T.W. Brown, J.P. Bumpus, H. Chen, M.B.F. Hunzaker, J. Lee, M. Mann, F. Merhout and A. Volfovsky
2018 Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Bartlett, J. and R. Norrie
2015 *Immigration on Twitter: Understanding Public Attitudes Online*. Demos, London. Available at https://apo.org.au/sites/default/files/resource-files/2015-04/apo-nid54788.pdf.

Blinder, S. and L. Richards
2020 UK public opinion toward immigration: Overall attitudes and level of concern. The Migration Observatory briefing. Centre on Migration, Policy and Society (COMPAS), University of Oxford, United Kingdom, January.

Bruns, A. and Y.E. Liang
2012 Tools and methods for capturing twitter data during natural disasters. *First Monday*, 17(4).

Campan, A., T. Atnafu, T. Truta and J. Nolan
2018 Is data collection through Twitter streaming API useful for academic research? *IEEE International Conference on Big Data*, pp. 3638–3643.

Campbell Collaboration
n.d. Home page. Available at www.campbellcollaboration.org/ (accessed 12 November 2020).

Cheong, P.H., R. Edwards, H. Goulbourne and J. Solomos
2007 Immigration, social cohesion and social capital: A critical review. *Critical Social Policy*, 27(1):24–49.

Conover, M.D., J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini and F. Menczer
2011 Political polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):89–96.

** All hyperlinks were working at the time of writing this report.*

Dennison, J. and L. Dražanová

2018 *Public Attitudes on Migration: Rethinking How People Perceive Migration – An Analysis of Existing Opinion Polls in the Euro-Mediterranean Region*. Observatory of Public Attitudes to Migration, European Union. Available at https://cadmus.eui.eu/handle/1814/62348.

Dennison, J. and A. Geddes

2019 A rising tide? The salience of immigration and the rise of anti-immigration political parties in Western Europe. *The Political Quarterly*, 90(1):107–116.

Esipova, N., J. Ray, A. Pugliese, D. Tsabutashvili, F. Laczko and M. Rango

2015 *How the World Views Migration*. IOM, Geneva. Available at https://publications.iom.int/books/how-world-views-migration.

European Political Strategy Centre (EPSC)

2019 *10 Trends Shaping Migration*. European Commission. Available at https://op.europa.eu/s/oq7V.

Freire, Y. and E. Graells-Garrido

2019 Characterization of local attitudes toward immigration using social media. arXiv:1903.05072.

Goyder, J.

1986 Surveys on surveys: Limitations and potentialities. *Public Opinion Quarterly*, 50(1):27–41.

Greven, T.

2016 *The Rise of Right-wing Populism in Europe and the United States: A Comparative Perspective*. Friedrich Ebert Foundation, Washington, D.C., pp. 1–8. Available at https://library.fes.de/opus4/frontdoor/index/index/docId/44073.

Grigorieff, A., C. Roth and D. Ubfal

2020 Does information change attitudes toward immigrants? *Demography*, 57(3):1117–1143.

Grow, A., D. Perrotta, E. Del Fava, J. Cimentada, F. Rampazzo, S. Gil-Clavel, E. Zagheni, R.D. Flores, I. Ventura and I.G. Weber

2021 How reliable is Facebook's advertising data for use in social science research? Insights from a cross-national online survey. MPIDR Working Papers WP-2021-006. Max Planck Institute for Demographic Research, Rostock.

Hawelka, B., I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos and C. Ratti

2014 Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

Hong, S. and S.H. Kim

2016 Political polarization on Twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, 33(4):777–782.

Hootsuite and We Are Social

2019 Digital 2019: Global digital overview. 31 January. Available at https://datareportal.com/reports/digital-2019-global-digital-overview.

Hutto, C. and E. Gilbert

2014 VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

International Organization for Migration (IOM)

2017 Surveys measuring public opinion on migration. Global Migration Data Analysis Centre (GMDAC). Available at https://migrationdataportal.org/infographic/surveys-measuring-public-opinion-migration (accessed 12 November 2020).

Ipsos

2019 Attitudes towards immigration. Survey conducted on behalf of IMiX. London. Available at www.ipsos.com/sites/default/files/ct/news/documents/2019-03/public-attitudes-towards-immigration-survey-for-imix.pdf.

Larsen, M.E., T.W. Boonstra, P.J. Batterham, B. O'Dea, C. Paris and H. Christensen

2015 We feel: Mapping emotion on Twitter. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1246–1252.

Leetaru, K., S. Wang, G. Cao, A. Padmanabhan and E. Shook

2013 Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5).

Li, D.K.

2019 Texas man charged with killing 12 women, police look at 750 other deaths for possible links. *NBC News*, 17 May. Available at www.nbcnews.com/news/us-news/texas-man-who-allegedly-posed-maintenance-worker-charged-killing-seven-n1006436.

McCormick, T.H., H. Lee, N. Cesare, A. Shojaie and E.S. Spiro

2017 Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods & Research*, 46(3):390–421.

Morstatter, F., J. Pfeffer, H. Liu and K.M. Carley

2013 Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):400–408.

Nature

2020 Stop the coronavirus stigma now. 7 April. Available at www.nature.com/articles/d41586-020-01009-0.

Paul, M.J. and M. Dredze

2012 A model for mining public health topics from Twitter. *Health*, 11(16-16):1.

Penninx, R., D. Spencer and N. Van Hear

2008 Migration and integration in Europe: The state of research. Economic and Social Research Council, Swindon.

Rankin, J.

2020 Britain's EU citizens "at risk of discrimination" after Brexit, say MEPs. *The Guardian*, 15 January. Available at www.theguardian.com/politics/2020/jan/15/eu-citizens-in-uk-risk-discrimination-in-jobs-and-housing.

Rowe, F., M. Mahony, E. Graells-Garrido, M. Rango and N. Sievers

2021 Using Twitter to track immigration sentiment during early stages of the COVID-19 pandemic. *Data & Policy*, 3(e36).

Salganik, M.J.

2019 *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Schwartz, C., M. Simon, D. Hudson and J. van-Heerde-Hudson

2021 A populist paradox? How Brexit softened anti-immigrant attitudes. *British Journal of Political Science*, 51(3):1160–1180.

Smith, D.

2020 Trump impeachment: Trump's acquittal offers glimpse of America's imploding empire. *The Guardian*, 5 February. Available at www.theguardian.com/us-news/2020/feb/05/trump-acquittal-impeachment-trial-america-imploding-empire.

Terpstra, T., R.J.P. Stronkman, A. de Vries and G.L. Paradies

2012 Towards a realtime Twitter analysis during crises for operational crisis management. In: *9th Proceedings of the International Conference on Information Systems for Crisis Response and Management* (Rothkrantz, L.J.M., J. Ristvej and Z. Franco, eds.). Simon Fraser University, Vancouver.

The Guardian

2020 The Observer view on the government's immigration plans. Observer editorial, 23 February. Available at www.theguardian.com/commentisfree/2020/feb/23/observer-view-on-immigration.

Tornes, A.
2021 Enabling the future of academic research with the Twitter API. Twitter, 26 January. Available at https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api.

Twitter
n.d. Tutorials: Filtering tweets by location. Available at https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location (accessed 26 August 2022).

Vosoughi, S., D. Roy and S. Aral
2018 The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wang, H., D. Can, A. Kazemzadeh, F. Bar and S. Narayanan
2012 A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In: *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics (ACL), Stroudsburg, pp. 115–120.

Wikipedia
n.d.a COVID-19 pandemic cases. Available at https://en.wikipedia.org/wiki/COVID-19_pandemic_cases (accessed 12 November 2020).
n.d.b List of incidents of xenophobia and racism related to the COVID-19 pandemic. Available at https://en.wikipedia.org/wiki/List_of_incidents_of_xenophobia_and_racism_related_to_the_COVID-19_pandemic (accessed 12 November 2020).

ADDITIONAL READING

Gottfried, G. and A.P. Aslaksen
2017 *Shifting Ground: 8 Key Findings from a Longitudinal Study on Attitudes towards Immigration and Brexit*. Ipsos, London.

9

SOCIAL MEDIA

SAMPLING MIGRANTS AND DIASPORAS

THE SAMPLING OF MIGRANTS THROUGH ADVERTISEMENTS ON FACEBOOK AND INSTAGRAM

Steffen Pötzschke[1]

Introduction

The recruitment of migrants into surveys is a complicated endeavour. While several established techniques exist, they all have their shortcomings. This is especially apparent in projects that aim to study specific migrant populations in different countries simultaneously. Such cross-national research seeks to use comparable sampling techniques in all countries under observation. However, differences in administrative and technical infrastructure and, consequently, in the available sampling frames can be expected (Reichel and Morales, 2017). For example, population registers do not exist in all countries, and where they do, they might not contain all the variables needed to identify migrants. Furthermore, researchers are not always granted (unrestricted) access to this information (Careja and Andreß, 2018). Other list-based sampling procedures, such as onomastic sampling (Schnell et al., 2017), presuppose not only the existence of suitable lists (e.g. telephone books) but also the inclusion of the target population in them. Random-route sampling, another often-used technique, restricts the sampling to specific areas, is relatively cost-intensive since interviewers are needed "on the ground", and becomes increasingly ineffective the smaller a target population is, compared to the overall population (Salentin, 2014). While some established methods, such as population register-based sampling, in theory, allow for the realization of probability samples that might enable the generalization of findings, their implementation is rather complex, time- and resource-consuming, and, as mentioned, might not be possible on a cross-national level, depending on the targeted countries.

In light of these challenges, migration scholars are increasingly exploring opportunities provided by new data sources and Internet-based services to expand their methodological toolkit (Pötzschke and Rinken, forthcoming). One such avenue is the employment of social networking sites (SNSs) to sample migrants for survey research projects. This seems appropriate as the use of SNSs has seen rapid growth since the start of the twenty-first century. Facebook (2020a), for example, reported 2.5 billion monthly active users worldwide in December 2019, of whom 1.66 billion visited the network daily. These figures indicate that the use of SNSs is widespread. It stands to reason that this

[1] Steffen Pötzschke is Deputy Head of the Panel Team at the GESIS – Leibniz Institute for the Social Sciences. He is a corresponding member of the Institute for Migration Research and Intercultural Studies and holds a doctorate from Osnabrück University (Germany). He can be reached at steffen.poetzschke@gesis.org and https://orcid.org/0000-0003-1039-275X.

(and many other) SNS is used not just by sedentary but also by mobile populations. Indeed, previous findings show that migrants use new communication technologies to maintain social relations over long distances (Emmer et al., 2016; Oiarzabal, 2012; Sanchez et al., 2018).

This chapter presents a sampling approach that employs advertisements on Facebook and Instagram to reach specific target populations. This method is especially suitable for surveys of otherwise hard-to-reach migrant populations and in cross-national research. The presented evidence underlines the technique's effectiveness and cost-efficiency.

State of the art

The use of Facebook as a sampling tool in survey research has gained increasing attention over the last decade. The capability to display targeted advertisements to selected user groups has proven most effective in this regard.

This sampling method has already gained considerable popularity in the area of medical and health research. In a systematic literature review in this field, Whitaker et al. (2017) report a total of 35 studies published between 2012 and 2017, which used Facebook advertisements to recruit participants. In recent years, various authors employed the method to study migrants. Carlini and colleagues (2015), for instance, used advertisements on Facebook (and other electronic media) to recruit Brazilian immigrants in the United States of America for a survey on substance abuse. Pötzschke and Braun (2017) recruited Polish immigrants in Austria, Ireland, Switzerland and the United Kingdom through the same method. Likewise, the MOBILISE project (Ersanilli and van der Gaag, 2021) sampled migrants via SNSs on a cross-national level – namely, Argentines, Moroccans, Poles, and Ukrainians in Germany, Spain and the United Kingdom. Finally, in their recent feasibility study, Pötzschke and Weiß (2021) targeted German emigrants living in any country outside the European continent.

Recent studies have also shown that SNSs, such as Facebook, can be successfully used to sample other hard-to-reach populations, such as individuals in precarious working conditions (Schneider and Harknett, 2019) and members of the LGBTQ community (Kühne and Zindel, 2020). Likewise, SNSs allow the implementation of large-scale surveys on rapidly emerging and developing topics, such as the COVID-19 pandemic in 2020 (Grow et al., 2020).

While all of the above-mentioned studies posted advertisements on Facebook, the feasibility of using other SNSs for sampling purposes is also increasingly investigated. This is a logical next step in this line of research because Facebook is not available in all countries. Even where it is, it might not be used by individuals in different age cohorts, from different backgrounds, etc., to the same degree. Rocheva and colleagues (forthcoming), for example, point out that the Russian SNSs VK (previously called VKontakte) and Odnoklassniki have more users than Facebook in the Russian Federation and also in most countries of origin of its biggest immigrant communities. Consequently, the authors provide a thorough description of their experience in employing these networks as sampling tools. Ersanilli and van der Gaag (2021) likewise have explored the use of VK to recruit Ukrainian migrants. Instagram is another network that is increasingly included in sampling efforts. The use of this SNS is convenient, as corresponding advertisements can be launched directly through the advertisement manager of its parent company (Facebook, Inc.) too. This allows researchers to sample participants

through Facebook and Instagram in a coordinated way using a coherent targeting strategy. Studies that have recently followed this approach include those by Kühne and Zindel (2020) and Pötzschke and Weiß (2021).

Facebook and Instagram advertisements as a sampling mechanism

The following provides only a short introduction to the use of advertisements on Facebook and Instagram as a sampling mechanism.[2] As mentioned above, the sampling method discussed in this chapter utilizes the marketing tool Facebook Ads[3] by which Facebook, Inc., offers the possibility to place advertisements on its SNSs Facebook and Instagram. Besides survey sampling, this marketing tool – more precisely its application programming interface (API) – has in recent years been used by scholars to generate estimates of migration stocks and flows (e.g. Palotti et al., 2019; Spyratos et al., 2019; Zagheni et al., 2017).

Designing and setting up advertisements

The first step of implementing an advertisement campaign on Facebook and Instagram consists of deciding on a campaign objective; this means the user behaviour which displaying an ad should trigger. In the survey context, advertisements will usually be employed to generate clicks on a link that guides users to an externally hosted questionnaire. At this point, it should be highlighted that the described sampling approach does not use any personal information that users share on Facebook's platforms as a basis of analysis. Indeed, researchers employing this method will neither need nor gain access to users' profiles, communication or the information included in them. SNSs are only used to recruit participants, while all data are collected through externally hosted online surveys. This approach builds on a basic functionality provided and administered by Facebook, Inc., and is fully covered by its user agreement. Consequently, this method avoids ethical problems at times associated with the use of big data.

An important issue regarding the discussed advertisements is their placement, since only users who see an ad can read and (possibly) react to it. The first decision that needs to be taken in this regard is on which devices the advertisements should be displayed. On Facebook and Instagram, their delivery can be restricted to either computers or mobile devices or be allowed on both.

Second, the advertisement platform(s) of delivery need(s) to be selected. As of November 2020, the offered choices consist of Facebook, Instagram, Audience Network and Facebook Messenger. Obviously, the placement decision has consequences for the size of the advertisements' potential target group. The more platforms are selected, the higher the number of users that might see the advertisements. However, from a methodological point of view, the stimulus that different users receive to click on an advertisement should be as similar as possible to avoid biases in the resulting sample. For this reason, and partly because the number of available platforms increased only in recent years, most of the scientific studies mentioned above displayed their advertisements on Facebook

[2] For a more detailed description, see: Pötzschke and Braun, 2017.

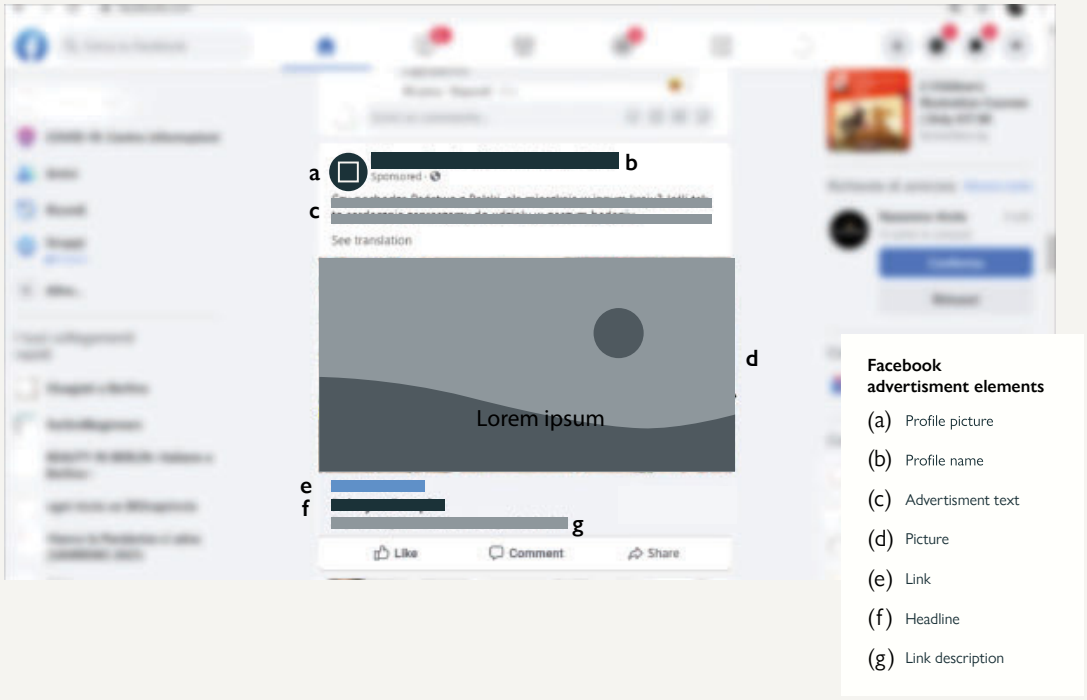
[3] More information is available at www.facebook.com/business/ads.

only. If additional platforms are used, measures should be taken to record the platform from which individual survey participants came. Pötzschke and Weiß (2021), for instance, used specific URL parameters (Facebook, 2020b) that were recorded in the data set.

The third placement decision concerns the sections in which the advertisements should be displayed on the chosen platforms. Several options – such as the Facebook News Feed, Instagram Feed, Facebook Marketplace, Messenger inbox and Messenger Stories – are available, depending on the choice in the previous step. Again, the aim of a consistent survey design should be considered when making this decision because of its implication for the visual design and presentation of the advertisements themselves. The most uncomplicated option to keep the stimulus for all individuals as constant as possible is the exclusive display of advertisements on news feeds. In this case, advertisements look nearly identical on computers and mobile devices. Additionally, this has the advantage of the advertisements being included between user-generated content (such as private posts and pictures), which promises great exposure. Furthermore, if both Facebook and Instagram are used, this option results in a high similarity of the advertisements in both environments. If additional placement options are selected, their respective user engagement should, again, be recorded in the data set.

Naturally, the design of individual advertisements should be carefully considered. Figure 1 shows their different components, using the design elements of Pötzschke and Braun (2017) as an example.

Figure 1. Elements of an advertisement to be displayed on the Facebook News Feed



Source: Pötzschke and Braun (2017), adapted. Copyright of the picture: Igor Mojzes.

The first two essential elements are the profile picture (a) and the profile name (b), which are displayed at the top of the ad. Both elements are taken directly from the Facebook account that is placing the advertisement. If the ad is created using the main account of the institution conducting a survey, the name of this account and its profile picture are used. However, a consequence of this is that the most prominent place remaining to display the survey title would be *underneath* the picture, the so-called headline (e). To circumvent this, the creation of a survey-specific Facebook page is recommended. Ideally, such a page should have the title of the survey as its name and a compact logo of the institution carrying out the project as the account picture, as shown in the example. The instalment of a specific survey page on Facebook also has the advantage of it being used to communicate further information regarding the project, reassuring potential participants of its legitimacy.

The text space offered by Facebook ads is very limited. Ideally, users should be able to read the whole main advertisement text (c) without clicking on a "See more" link. This means, however, that only approximately 125 characters can be used to convey a message in this primary text element. The wording of each text element in the advertisement is important and should directly motivate eligible users to participate in the survey.

Besides the text, pictures are a vital component of Facebook ads as the network itself is a very visual medium. This applies even more to Instagram. Therefore, the decision on a picture or a group of pictures should not be made lightly.[4] Pictures serve to capture users' attention and should therefore relate to the target population's country or culture of origin and, if possible, the survey's topic. Ramo and Prochaska (2012) showed that otherwise identical ads using different pictures varied substantially regarding the number of clicks they produced. On the one hand, this can be explained by the fact that certain pictures are likely to appeal more to the target population than others. On the other hand, it is probable that different strata of the target population would vary in their respective preferences. As these effects can hardly be quantified before a study without extensive pretests, it is advisable to use several pictures for any given advertisement campaign.

Specifying the target group and the importance of using several ad sets

It is an essential feature of advertisements on Facebook and Instagram that they can be used to target specific subpopulations among their users. Consequently, defining the target population is a crucial aspect of this procedure. The Facebook Ads Manager offers a range of indicators that are relevant in this regard. Besides age and gender, the most important variables for migration research purposes are physical location, language use, and indicators that are assigned to users who do not live in their country of origin. Physical locations can, for example, be specified as continents, countries, cities, or just radiuses around addresses or points of interest. There are options that indicate how users should relate to these locations. Surveys will usually target users who live in the chosen area, and thereby define the sampling's geographic scope. Through the language variable, the targeting can be focused on users who speak a language that is distinct from the main language of their place of residence. Two indicators are directly related to users' mobility and migration experiences. The first allows the targeting algorithm to select users who previously lived in specific countries (e.g. "lived in Argentina"). While this variable is offered for many countries, it is not available for all of them. The second variable is more generic and broadly identifies individuals who do not live in their home country (i.e. "lives

[4] It is also possible to use videos instead of static pictures. However, to date, most scientific studies employ pictures, and no systematic information is available regarding possible commonalities and differences in the performance of advertisements featuring either media type.

abroad"). To reach migrants from countries that cannot be targeted directly, the latter variable might be used in combination with other criteria, such as specific interests and educational facilities visited. Practical examples of such targeting are provided by Ersanilli and van der Gaag (2021).

When designing an advertisement campaign to sample migrants or other target groups through Facebook and Instagram, the structure of such campaigns should be taken into account. A campaign consists of one or more ad sets that feature one or more advertisements. The aforementioned targeting criteria are defined at the ad set level. Importantly, Facebook has implemented an algorithm that pushes those ads within a given ad set that perform the best with regard to the specified campaign objective. Hence, if the objective is to generate clicks to a website, Facebook will increasingly display those ads which receive the most clicks over the lifetime of a campaign. As Arcia (2014) points out, this can lead to a biased sample in a setting that employs several advertisements featuring different pictures, if users who share certain traits (e.g. of a socioeconomic or cultural nature) are more inclined to click on Facebook ads than others, and if they tend to prefer a specific picture. The same logic applies to the target population's geographic distribution: if an ad set targets various countries of residence, the Facebook algorithm will not automatically consider these relevant categories to stratify the sample. This means that if significantly more members of the target population live in country A than in country B, while the likelihood that people in either of these countries will click on the ad is identical, the algorithm is likely to increasingly display the advertisement in country A, as the ad will accumulate reactions more rapidly this way. Consequently, it is advisable to include the geographic aspect in the planning of ad sets, if relevant differences are expected. To evaluate individual ad sets' performance, researchers should make sure that the ad set through which a specific participant is recruited is recorded in the data set. Again, the use of URL parameters is the easiest way to do so.

Evidence of cost-effectiveness and billing mechanism

While the existing evidence suggests that the costs for sampling through Facebook (and Instagram) vary, this method is indeed very cost-effective compared to more established ones. In their systematic review, Whitaker et al. (2017) reported mean costs of approximately EUR 12.14 (USD 14.41)[5] per completed questionnaire. However, these authors compared health-related studies which mostly targeted very specific subpopulations and focused on rather sensitive topics. This might partly explain why studies in other fields usually reported much lower costs: in the survey of hourly workers by Schneider and Harknett (2019), the average recruitment cost amounted to approximately EUR 4.85 (USD 5.76, own calculation) per questionnaire completed by individuals in the analytical sample. For their large cross-national COVID-19 survey, Grow et al. (2020) put the average cost per completed questionnaire at EUR 1.05. Rosenzweig and colleagues (2020) reported average costs of approximately EUR 0.13 (USD 0.16) per completed questionnaire in their survey of members of the general Mexican population, and approximately EUR 0.75 (USD 0.89) for a similar undertaking in Kenya. The costs in the cross-national study of Polish migrants by Pötzschke and Braun (2017) amounted on average to EUR 0.47 per questionnaire completed by eligible participants, while Ersanilli and van der Gaag (2021) reported costs of EUR 1.45 in their survey of migrants of the same origin (in partly different countries of residence).

[5] The exchange rate used is USD 1 = EUR 0.8426 (European Central Bank). More information is available at www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/eurofxref-graph-usd.en.html (accessed 16 November 2020).

Advertisements on Facebook and Instagram do not have a fixed price; instead, the costs are determined by an auctioning procedure. Once an advertisement is activated,[6] the buyer engages in an automated bidding competition with other advertisers who target the same user group. This means that sampling cost for a given project and target population can only be roughly estimated beforehand, using extant survey literature as orientation. However, it should be taken into consideration that the effectiveness of the recruitment – and consequently the incurred costs – depends on a variety of factors, such as the sociodemographic composition of the target population, the design of the advertisements, and the topic of the survey and its appeal to the members of the target population.

While sampling costs cannot be precisely calculated beforehand, the Ads Manager allows users to specify the available budget: specific budgets can be assigned to individual ad sets, and an additional spending limit can be defined at the campaign level.

Main limitations and advantages

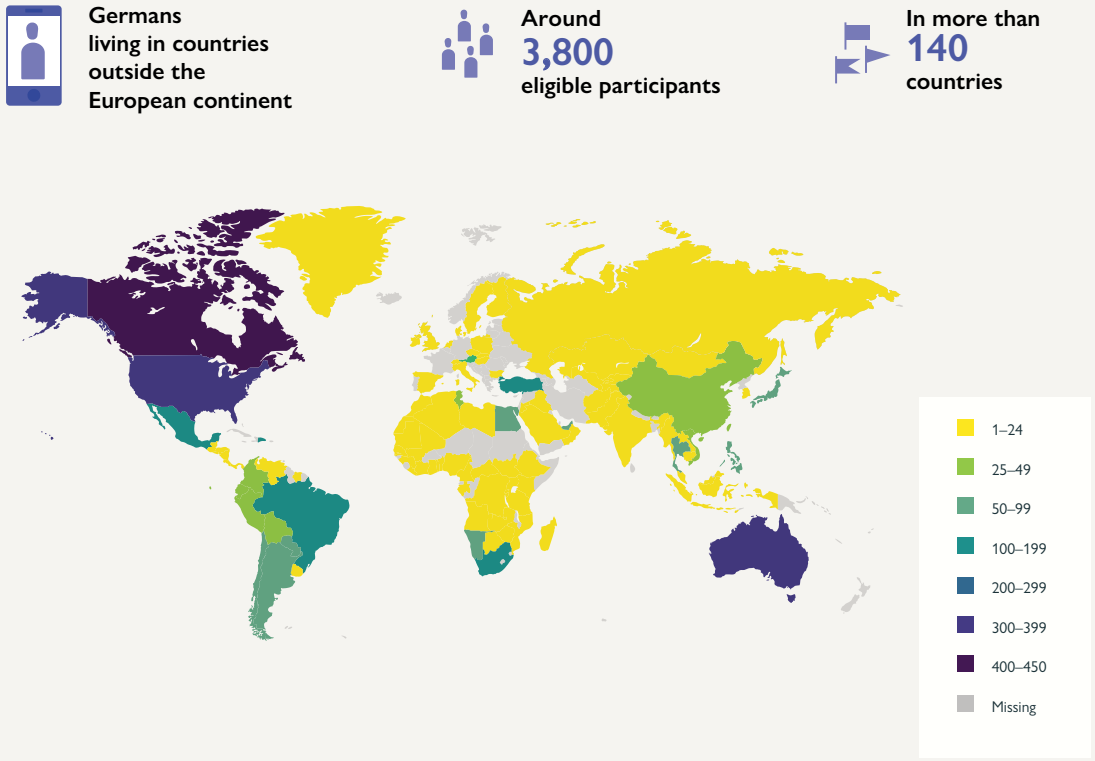
The first limitation of the presented sampling approach is that Facebook and Instagram cannot be accessed in all countries. This means that residents of some countries cannot be targeted through these advertisements. Furthermore, even where these SNSs are available in principle, their use is dependent on a certain infrastructure (e.g. Internet connection, functioning power grid). A rather obvious limitation is that a sample recruited through advertisements in an SNS will, in most cases, exclusively consist of individuals who are its registered users. However, researchers might try to broaden this scope by using an additional snowball element (cf. Pötzschke and Weiß, 2021). It should also be considered that people in different age cohorts might use Facebook to varying degrees, though the simultaneous targeting of potential participants through Instagram might be a way to address this issue to a certain extent. Finally, migrants from some countries cannot be targeted as directly through the Facebook Ads Manager as others because the indicator identifying migrants of specific national origin does not exist for some States (e.g. Syrian Arab Republic). The lack of detailed information on how the targeting variables in the Facebook Ads Manager are constructed adds an additional layer of uncertainty and potential bias.

Nevertheless, the technique also has clear advantages. It is especially helpful in cases where traditional sampling frames are not available or if the implementation of alternative sampling strategies is not feasible due to financial considerations. The approach is particularly suitable for surveying small and otherwise hard-to-reach migrant populations, and for countries in which established sampling strategies might be difficult to implement. Compared to more conventional sampling approaches, the described technique is very cost-effective. Furthermore, it can be implemented on short notice, making it ideal for researching the impact of current events on migrants and migration. While this approach makes use of software and information owned by a private company, it is not dependent on direct cooperation with this entity. This means that it is also feasible for institutions and researchers of all career stages who lack personal or institutional connections that would allow them to collaborate directly with big cooperations such as Facebook, Inc. Regarding cross-national research, it is advantageous that sampling through Facebook and Instagram can be implemented in a wide range of countries. Besides immigration studies, the approach seems well suited to advance research

[6] Advertisements are not fielded directly after activation but undergo an automated review process first.

on emigrant communities and diasporas because it does not force researchers to focus on a limited number of countries merely for logistical or financial reasons. To illustrate this potential, Figure 2 shows the distribution of the sample achieved by the German Emigrants Overseas Online Survey using a total advertisement budget of less than EUR 2,400 (Pötzschke and Weiß, 2021).

Figure 2. Geographic distribution of participants in the German Emigrants Overseas Online Survey



Source: Pötzschke and Weiß, 2021.

Note: This map is for illustration purposes only. The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by the International Organization for Migration.

The mentioned project targeted Germans living in countries outside the European continent and collected data from approximately 3,800 eligible participants[7] from more than 140 countries and territories. These results are encouraging as they stress that sampling through SNSs might indeed be a helpful addition to the toolkit of scholars and organizations interested in studying specific emigrant populations globally.

Conclusion

Not only are social networking sites repositories of large data sets, but they can also be used effectively as survey sampling tools. This approach is especially helpful for surveys targeting migrants and other hard-to-reach populations both within individual countries and in cross-national research. In the latter context, the method can be particularly beneficial as it allows the implementation of a coherent sampling strategy in most countries worldwide. The available evidence suggests that

[7] The target population are people born in Germany or in possession of German citizenship – or both – living in any country other than Germany. Germans living in European countries were not directly targeted by the ad campaign but allowed to participate in the survey.

the technique is very cost-effective; however, its efficiency might vary depending on the population and countries under research. More generally, when considering using this sampling method, its nonprobability character and the resulting implications for data analysis should be considered. In short, this sampling approach seems well suited in scenarios in which: (a) it complements other sampling strategies to counterbalance known weaknesses (e.g. to sample individuals that are not listed in telephone registries); (b) it is employed as the sole strategy in the absence of other feasible approaches to sample hard-to-reach populations (e.g. if population registers are not available or if the target population is not identifiable in them); and (c) if the analytical goal does not demand the generalization of findings to broader population groups. In the context of the current COVID-19 pandemic or similar future events, sampling via SNSs could be particularly advantageous because it does not involve face-to-face interactions between project staff and participants, and it can easily be employed for self-administered web surveys or to recruit participants for qualitative voice over Internet Protocol (VoIP) interviews.

REFERENCES*

Arcia, A.

2014 Facebook advertisements for inexpensive participant recruitment among women in early pregnancy. *Health Education & Behavior*, 41(3):237–241. Available at https://doi.org/10.1177/1090198113504414.

Careja, R. and H.J. Andreß

2018 In search of a frame: Challenges and opportunities for sampling immigrant minorities. *Comparative Migration Studies*, 6(1):37. Available at https://doi.org/10.1186/s40878-018-0103-5.

Carlini, B.H., L. Safioti, T.C. Rue and L. Miles

2015 Using Internet to recruit immigrants with language and culture barriers for tobacco and alcohol use screening: A study among Brazilians. *Journal of Immigrant and Minority Health*, 17(2):553–560. Available at https://doi.org/10.1007/s10903-013-9934-1.

Emmer, M., C. Richter and M. Kunst

2016 *Flucht 2.0: Mediennutzung durch Flüchtlinge vor, während und nach der Flucht*. Institut für Publizistik- und Kommunikationswissenschaft, Freie Universität Berlin. Available at www.polsoz.fu-berlin.de/kommwiss/arbeitsstellen/internationale_kommunikation/Media/Flucht-2_0.pdf.

Ersanilli, E. and M. van der Gaag

2021 Data report: Online surveys. Wave 1. MOBILISE working papers. Available at https://doi.org/10.31235/osf.io/79gca.

Facebook

2020a Form 10-K: Annual Report 2019 – Facebook, Inc. Available at http://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/45290cc0-656d-4a88-a2f3-147c8de86506.pdf.

2020b Add URL parameters to your ads. Facebook for Business – Business Help Centre. Available at www.facebook.com/business/help/1016122818401732.

Grow, A., D. Perrotta, E. Del Fava, J. Cimentada, F. Rampazzo, S. Gil-Clavel and E. Zagheni

2020 Addressing public health emergencies via Facebook surveys: Advantages, challenges, and practical considerations. *Journal of Medical Internet Research*, 22(12):e20653. Available at https://doi.org/10.2196/20653.

Kühne, S. and Z. Zindel

2020 Using Facebook and Instagram to recruit web survey participants: A step-by-step guide and application. *Survey Methods: Insights from the Field*. Available at https://doi.org/10.13094/SMIF-2020-00017.

** All hyperlinks were working at the time of writing this report.*

Oiarzabal, P.J.
2012 Diaspora Basques and online social networks: An analysis of users of Basque institutional diaspora groups on Facebook. *Journal of Ethnic and Migration Studies*, 38(9):1469–1485. Available at https://doi.org/10.1080/1369183X.2012.698216.

Palotti, J., N. Adler, A.J. Morales, J. Villaveces, V. Sekara, M. Garcia Herranz, M. Al-Asad and I. Weber
2019 Real-time monitoring of the Venezuelan exodus through Facebook's advertising platform. Qatar Computing Research Institute, United Nations Children's Fund, Massachusetts Institute of Technology, iMMAP and Global Protection Cluster. Available at https://data2.unhcr.org/en/documents/details/68638.

Pötzschke, S. and M. Braun
2017 Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. *Social Science Computer Review*, 35(5):633–653. Available at https://doi.org/10.1177/0894439316666262.

Pötzschke, S. and S. Rinken (eds.)
Migration Research in a Digitized World: Using Innovative Technology to Tackle Methodological Challenges. Springer (forthcoming).

Pötzschke, S. and B. Weiß
2021 Realizing a Global Survey of Emigrants through Facebook and Instagram. GESIS – Leibniz Institute for the Social Sciences. Available at https://doi.org/10.31219/osf.io/y36vr.

Ramo, D.E. and J.J. Prochaska
2012 Broad reach and targeted recruitment using Facebook for an online survey of young adult substance use. *Journal of Medical Internet Research*, 14(1):e28. Available at https://doi.org/10.2196/jmir.1878.

Reichel, D. and L. Morales
2017 Surveying immigrants without sampling frames: Evaluating the success of alternative field methods. *Comparative Migration Studies*, 5(1):1. Available at https://doi.org/10.1186/s40878-016-0044-9.

Rocheva, A., E. Varshaver and N. Ivanova
Targeting on social networking sites (SNS) as a sampling strategy for online migrant surveys: The challenge of biases and search for possible solutions. In: *Migration Research in a Digitized World: Using Innovative Technology to Tackle Methodological Challenges* (Pötzschke, S. and S. Rinken, eds.). Springer (forthcoming).

Rosenzweig, L.R., P. Bergquist, K. Hoffmann Pham, F. Rampazzo and M. Mildenberger
2020 Survey sampling in the Global South using Facebook advertisements [preprint]. SocArXiv. Available at https://doi.org/10.31235/osf.io/dka8f.

Salentin, K.
2014 Sampling the ethnic minority population in Germany: The background to "migration background". *Methods, Data, Analyses*, 8(1):25–52. Available at https://doi.org/10.12758/mda.2014.002.

Sanchez, G., R. Hoxhaj, S. Nardin, A. Geddes, L. Achilli and R.S. Kalantaryan
2018 *A Study of the Communication Channels Used by Migrants and Asylum Seekers in Italy, with a Particular Focus on Online and Social Media.* European Commission, Brussels. Available at http://hdl.handle.net/1814/61086.

Schneider, D. and K. Harknett
2019 Consequences of routine work-schedule instability for worker health and well-being. *American Sociological Review*, 84(1):82–114. Available at https://doi.org/10.1177/0003122418823184.

Schnell, R., T. Gramlich, T. Bachteler, J. Reiher, M. Trappmann, M. Smid and I. Becher
2017 A new name-based sampling method for migrants. *Methods, Data, Analyses*, 7(1). Available at https://doi.org/10.12758/mda.2013.001.

Spyratos, S., M. Vespe, F. Natale, I. Weber, E. Zagheni and M. Rango
2019 Quantifying international human mobility patterns using Facebook Network data. *PLOS ONE*, 14(10). Available at https://doi.org/10.1371/journal.pone.0224134.

Whitaker, C., S. Stevelink and N. Fear
2017 The use of Facebook in recruiting participants for health research purposes: A systematic review. *Journal for Medical Internet Research*, 19(8):e290. Available at https://doi.org/10.2196/jmir.7071.

Zagheni, E., I. Weber and K. Gummadi
2017 Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43(4):721–734. Available at https://doi.org/10.1111/padr.12102.

10

OTHER

SOCIAL MEDIA

SAMPLING MIGRANTS AND DIASPORAS

USING GOOGLE ANALYTICS AND ONOMASTIC ANALYSIS IN DIASPORA MAPPING

Michael Newson[1] *and Niklas Sievers*[2]

Introduction

An increasing number of governments throughout the world have come to recognize the potential and opportunities for engaging their diaspora populations in order to stimulate social and economic development in their countries of origin – be it through philanthropy, skills/knowledge transfer, investment or trade (Diaspora Liaison Office of Zambia, 2011; IOM, 2010; Schwartz, 2007). Diaspora mapping, which aims to provide a snapshot of the size, geography and characteristics of diaspora populations – meaning, individuals with historical, cultural, religious and affectional ties to countries of origin – is a critical first step in any diaspora engagement strategy (IOM and MPI, 2012; IOM, 2019). While diaspora mapping still relies primarily on traditional data – mainly population censuses, residence registration procedures and immigration statistics from destination countries (UNECE, 2012) – big data analysis may be leveraged to fill in information gaps and provide new insights into the characteristics of diaspora populations (Alinejad et al., 2019; *Routed Magazine* and iDiaspora, 2021; IOM, 2021). Using web analytics and onomastic analysis, IOM has been experimenting with new data sets to better understand and map diaspora populations.

Objectives of diaspora mapping

Diaspora mapping projects vary significantly depending on their size (mapping the diaspora in one country or a small number of countries or globally), budget and purpose. In most cases, governments are seeking a broader understanding of the diaspora at large, including where they are located, whether there have been specific "waves" or trends in migration patterns, how and the extent to which they engage with the country of origin, etc. (EUDiF, n.d.; IOM and MPI, 2012). Governments can then use this information for a variety of reasons, including considering resource allocation in embassies and consulates, understanding communications touchpoints and developing a broad communications

[1] Michael Newson is Senior Labour Mobility and Social Inclusion Specialist at IOM. He holds an MBA from the University of Warwick, a master's from York University (Canada) and a bachelor's degree from the University of British Columbia. His fields of expertise are labour migration, diaspora engagement and migrant integration.

[2] Niklas Sievers is Data Knowledge Officer at IOM GMDAC. He developed several research projects exploring the potential and pitfalls of innovative data sources, methods, and tools for migration policy and research. Before joining the United Nations System, he worked as an Advisor in the European Union Advisory Unit of PricewaterhouseCoopers and as a Lecturer at Humboldt University. He holds an MSc from the London School of Economics and Political Science in the social sciences and two bachelor's degrees from Leuphana University.

strategy, and developing policies and programmes that aim to better leverage the types of engagement that diaspora populations are already involved in with their communities of origin (iDiaspora et al., 2021; IOM, 2021).

In other cases, the objectives are much more specific. Governments may be seeking highly skilled individuals with expertise within a specific sector (health or information and communications technology, for example), diaspora business owners who may be able to stimulate trade, or high net worth diaspora members who may be interested in investing. The information generated through these types of diaspora mappings can then be used for various purposes, including to (a) develop targeted communications strategies for diaspora engagement in specific projects or initiatives, (b) inform policies to increase competitiveness for talent within specific sectors, or (c) develop a database of diaspora talent that can be called on for various types of requests or assignments in the future (IOM et al., 2021).

Ultimately, the design and methodology for a diaspora mapping project will depend on the broader objective of the diaspora engagement to which the mapping is linked. Regardless of the purposes, however, each project comes up against a common set of challenges in the availability and collection of data and information (Fabos et al., 2021). When it comes to general mapping projects, the most common approach involves analysing United Nations and government migration data combined with a small-scale survey (either online or in person), key informant interviews and focus group discussions. While migration data from the United Nations Department of Economic and Social Affairs can provide a picture of the migrant stock in a country, it is limited to the country level, without providing an understanding of how/where that population may be distributed, nor does it reveal much about changes in migration flow. Although immigration data from countries of destination can provide more detailed data, providing a clearer picture of both stock and flow from different countries of origin, it still is not able to provide greater detail on the distribution of the diaspora within the country and does not account for irregular migrants who may be a considerable share of the population in some cases. Census data may help to address these shortcomings in some cases. Still, there is a question of timeliness and accessibility, as finding, going through and deciphering data captured in various languages can be labour-intensive and time-consuming for researchers charged with diaspora mapping (Taylor et al., 2014).

In regard to the more qualitative data intended to provide a sense of diaspora character, needs/interests, and existing engagement, these tools also have their challenges as they tend to capture those diaspora members who are most engaged and interested in relations with the country of origin, rather than reflecting the broader diaspora community at large. Key informant interviews and focus group discussions will tend to capture data from leaders of diaspora associations or those active within the diaspora network in the country of destination. In contrast, online surveys will tend to be completed by those who have a particular interest in engagement already. Thus, the results usually cannot be said to provide a good picture of the broader diaspora population. When it comes to more targeted mapping, the most common approaches would be to coordinate communications through existing more specialized diaspora associations – specific professional diaspora associations related to the medical profession or information and communications technology, for example, or diaspora chambers of commerce – depending upon the particular interest for engagement (IOM, 2022).

This method of mapping creates two key challenges: (a) it only captures those who are well networked and engaged in the diaspora and thus may miss out on diaspora members with relevant skills and interests but who are less engaged with the diaspora community; (b) when establishing proactive outreach, inevitably you reach out to a broader range of diaspora members with a broader set of skills and knowledge than you are able to engage within the short or medium term meaningfully. As a result, the servers and hard drives of government diaspora units are littered with unused and underused databases containing the professional backgrounds and contact information of thousands of diaspora members who eagerly signed up to see how they could support development in their countries of origin. The lack of meaningful engagement following the initial outreach inevitably creates disappointment and frustration among diaspora members who may already have had little faith in the governments of their countries of origin. Thus, it is vital that diaspora engagement is framed in a mutually beneficial context and nurtures long-term links (Gevorkyan, 2019). Otherwise, even when done in good faith and with the intention of follow-up, it may risk creating frustration and distrust between the government and the skilled diaspora members with whom they wish to connect.

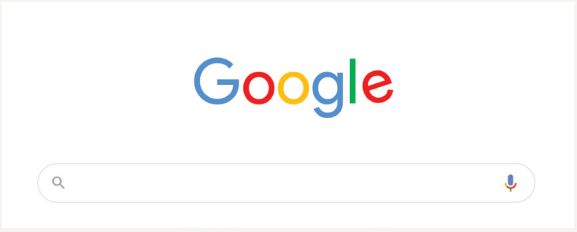
Using innovative data analysis methods in diaspora mapping

As a complement to the use of traditional data, in the last decade, several initiatives have leveraged big data sources to fill in information gaps and provide new insights in analyses of diaspora communities. These approaches drew on a range of innovative data sources and methods, including social media, Google Trends data, and onomastic analyses. One of the earlier initiatives has investigated the distribution of the diaspora of the Basque Country by analysing Facebook groups (Smith, 2000). After the rise of Twitter, Jones et al. (2011) used data from the social media service to track popular hashtags surrounding relevant events for diaspora members in their countries of origin (Demilt, n.d.). The potential of Google Trends data to support diaspora mapping became evident over time in several exploratory initiatives. Williams and Ralphs (2013) pointed out that Google searches for "Polski" were closely linked to the statistical figures on Polish nationals in the United Kingdom. Böhme et al. (2020) demonstrated in their paper "Searching for a better life" that Google Trends data can contribute to projecting international migration movements based on the frequency of search hits for a defined set of migration-related words, such as passport, embassy and asylum. In 2020, the United Nations Development Programme in Serbia replicated a similar approach and mapped the Serbian diaspora by analysing Serbian search requests on Google in countries with sizeable Serbian diaspora communities, such as Germany and Austria (Drašković, 2020). These approaches of deploying social media and Google data in diaspora mapping are supported by innovative data analysis methods, such as onomastic analysis. For instance, IOM has conducted onomastic analysis – meaning, the study of names and their origins – of United Kingdom and Russian business databases to identify members of Georgian, Armenian and Azerbaijani diasporas (UNECE, 2012). Further, this approach can support identifying neighbourhoods with a high concentration of diaspora members, which can be considered for specific diaspora engagement programmes through dedicated outreach material and information (IOM, 2017).

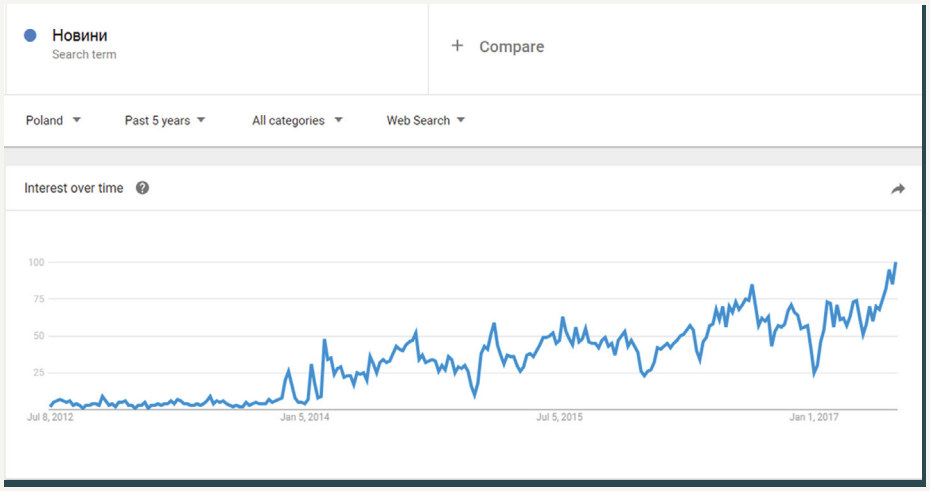
Exploring the use of Google search data and web traffic

Our first insight in using web traffic as a means of diaspora mapping recognized that people's Internet searches might indicate their nationality or country of origin. Using the term "balikbayan", we found that searches for this term aligned closely with the size of the Filipino population in each country and within each of the states when we zoomed into searches from the United States of America. Then, in 2017, there was increasing news about the rapid growth of labour migration from Ukraine to Poland, but figures were difficult to come by. In tracking the growth of searches for the Ukrainian term for "news" (новини) coming from Poland, we thought we could use this growth as a proxy to measure the growth of the Ukrainian population, as well as the seasonality of migration. The data did indicate steady and rapid growth in searches as well as sharp declines each year at the end of December/ beginning of January, a time when much of the Ukrainian population would return home for holidays.

Figure 1. Google searches in Poland for the Ukrainian word for "news"



In tracking the growth of searches for the Ukrainian term for "news" (новини) coming from Poland, this growth could have been used as a proxy to measure the growth of the Ukrainian population, as well as the seasonality of migration.



From there, we realized that web searches and website visits could indicate a person's nationality/ country of origin and that this could provide us with richer, more detailed data. National media/ news websites will tend to be accessed almost exclusively by persons from that specific country (as evidenced by the fact that the vast majority of views of those sites come from within the country itself). We could thus assume that the vast majority of views of these types of sites coming from outside the given country are likely coming from diaspora/emigrant communities. Therefore, by accessing the Google Analytics pages of these national media/news websites, we could map out the

concentrations and characteristics of the diaspora populations. We tested this using three Moldovan media sites and compared viewership from different cities in Italy and Canada; the similarity in results across all three websites increased our confidence in our earlier assumptions. Google Analytics can also provide additional information about viewers to these websites, including, for example, age group, gender and "interest groups" at the country or even the city level.

Figure 2. Percentage of website views from Italy coming from different Italian cities for three Moldovan media sites

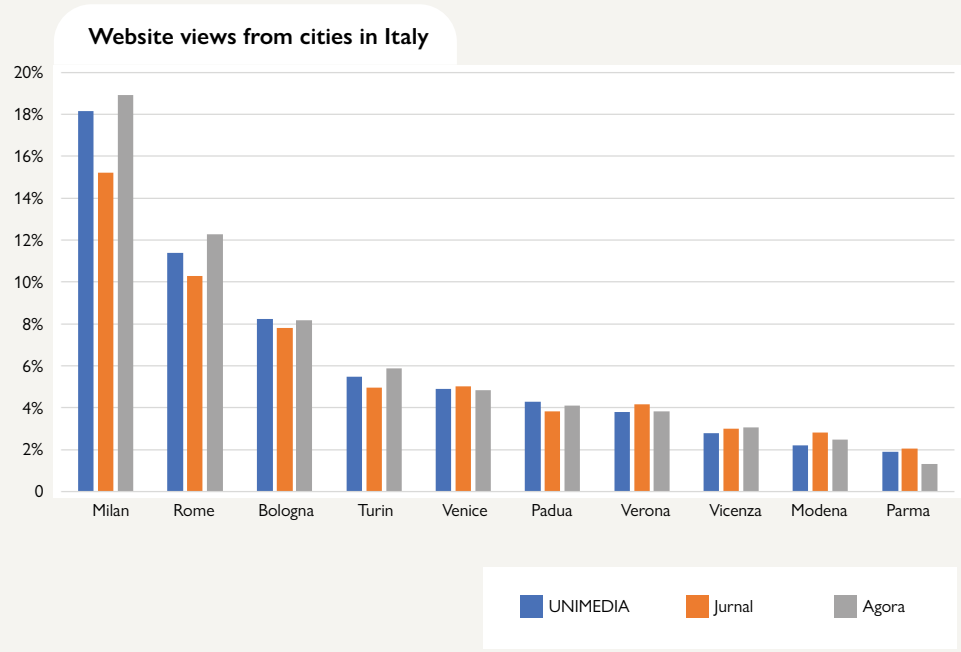
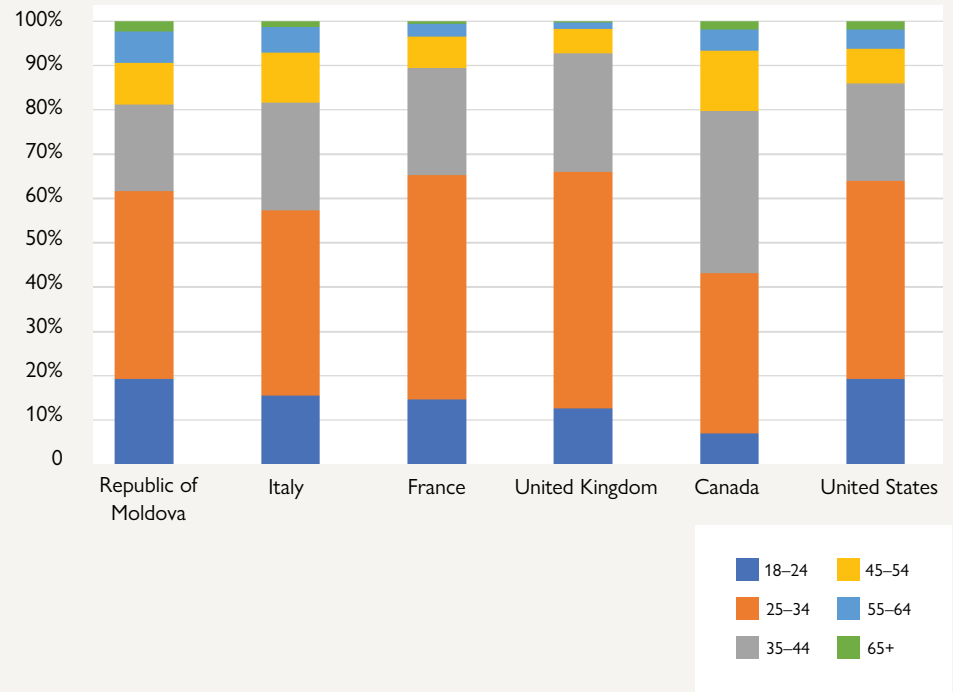


Figure 3. UNIMEDIA viewers by age group



While this mapping method cannot provide one with information in terms of the total numbers of the population in a country or city, it can provide excellent and up-to-date estimates of where diaspora communities are concentrated within any given country and across the world. Additionally, we can identify specific stories or types of news/information that are of particular interest to different diaspora groups, allowing stakeholders to develop more informed and targeted outreach and communications strategies for different segments of the diaspora. In tracking change in viewership of sites over time (assuming overall populations remain relatively stable), we can track the interest of diaspora communities in different events in the country of origin. In the Republic of Moldova, for example, the elections in 2020 resulted in a massive spike in readership from Italy, demonstrating the political engagement of the Moldovan diaspora in Italy in comparison to other countries of destination where increases were more muted.

Since first piloting this method of diaspora mapping in the Republic of Moldova, we have also applied it to Armenian websites, with similarly revealing results. We are now looking at how this initial work can be improved upon and mainstreamed into IOM's diaspora mapping activities.

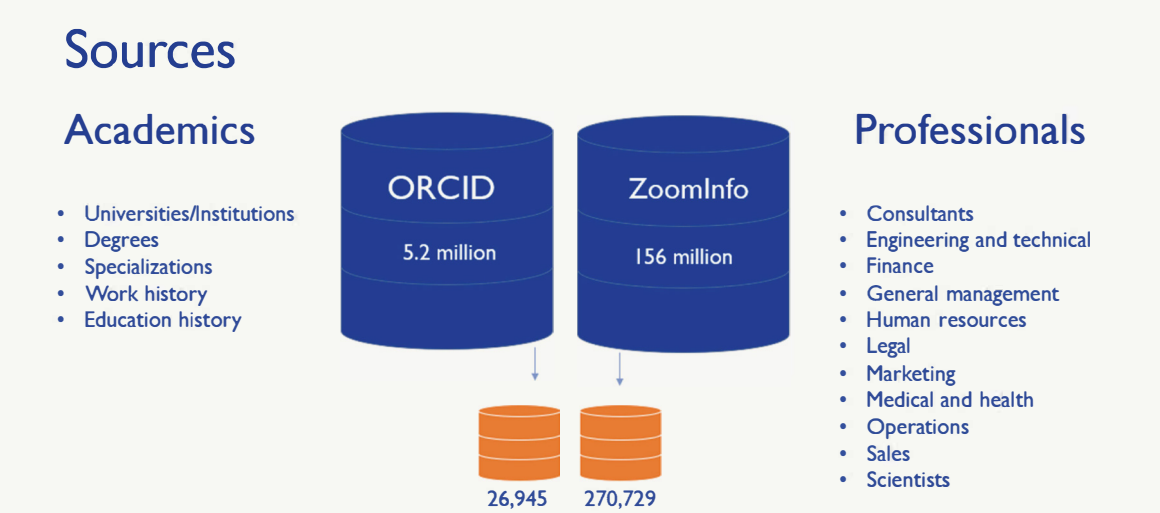
Applying onomastic analysis in diaspora mapping

Onomastics is the study of the origin of names, names (both first and family names) often being a good indicator of one's ethnic origin and country of origin or ancestry. On this front, IOM has been working with Namsor[3] using their application programme interface (API) to identify diaspora members within large public databases with the skills, knowledge and expertise sought by stakeholders within the country of origin. The databases to be analysed will vary depending on the nature of the diaspora and the types of profiles being sought within the mapping exercise. If the purpose is trade and investment, business registries within countries of destination may be analysed. If the objective is skills and knowledge transfer, the Open Researcher and Contributor ID (ORCID) database or others may be used. If the purpose is more general, LinkedIn can be mined for names of persons from a particular country of origin.

IOM has piloted this approach in both Armenia and Georgia to date, with promising results. In Armenia, for example, the research allowed IOM's office to identify medical professionals among the Armenian diaspora in France and the United States, with experience in treating patients with COVID-19. Linking these diaspora members via videoconference with health officials in Yerevan to share knowledge and experience, IOM was able to support the Armenian Government's preparedness and response to the pandemic. In Georgia, mining data from the United Kingdom Companies House and similar such registries in other countries of destination, IOM has been able to identify hundreds of Georgian diaspora business owners and CEOs. In January 2021, over 200 of these diaspora members participated in a virtual investment and business networking fair linking Georgian diaspora business owners with businesses in Georgia seeking trade opportunities and investors.

[3] Available at https://namsor.app/.

Figure 4. Results of onomastic analysis of Armenian names in the Open Researcher and Contributor ID (ORCID) and ZoomInfo databases



This method of diaspora mapping avoids several of the challenges listed above. First of all, onomastic analysis allows for capturing information on a large number of diaspora members, not just those that are active and known within the diaspora community. The net can be cast much wider and individuals with unique skills more easily identified within the diaspora. This is beneficial not only when you are trying to identify as many persons as possible but also when you are looking to identify individuals with very specific skill sets. For example, for a project in Georgia, IOM was seeking diaspora members with expertise in plant sciences. Using onomastic analysis, the office was able to identify over 20 Georgian diaspora members with this skill set – something that would have been near impossible using traditional mapping methods. This method also allows the mapping process to be more tailored to the specific interests and priorities of the study. Databases can be selected according to specific sectors or occupations as well as specific countries of destination depending on the focus of the diaspora mapping exercise.

Furthermore, as capturing this data does not require any initial engagement or effort from the diaspora members themselves, this approach avoids the potential frustration and resentment of diaspora members who would previously register themselves on a database and then see no follow-up. However, the flip side to this advantage are the challenges and low response rates that may occur when cold-calling members of the diaspora to enquire about their interest in particular initiatives. Indeed, as with all forms of big data analysis, but particularly in cases such as these where personal information is being captured, it is critical that private data is protected and its capture/storage avoided whenever possible. In pursuing this approach, IOM has worked carefully to ensure that it follows the Organization's data protection principles.

It should also be noted that onomastic analysis may not be an appropriate way of capturing data on the diaspora in all circumstances. In cases where first and last names tend to be similar across a wide number of countries – in Latin America or many anglophone countries, for example – onomastic analysis may not be sufficiently accurate to identify countries of origin based on names reliably. However, there may be other indicators (such as education history, for example) that could potentially serve as proxies to identify particular diaspora members within professional databases.

Conclusions

Through our piloting of web analytics and onomastic analysis in diaspora mapping over the past three years, we have seen the impressive potential of these new tools to fill in some of the gaps and shortcomings of traditional diaspora mapping techniques. These new big data methods can serve as a key enhancement – but not a replacement – for IOM's traditional practices in diaspora mapping to add new layers of understanding and provide actionable data and information to be used in policy development, communications strategies and programme development.

Reflecting on the use of web analytics, while it certainly cannot replace traditional research methods for diaspora mappings, it can serve as a powerful new tool that provides new additional revelations and nuances to existing research methods. In particular, it can provide more detailed information on where diaspora communities are concentrated (and to what extent) within countries of destination; it can also potentially provide information on the growth of diaspora communities in different countries as migration and settlement patterns may change both in terms of countries of destination and specific cities of settlement. These changing migration patterns will likely be reflected in changing patterns of web traffic as well.

In terms of outlining the different characteristics and interests of the diaspora, web analytics may provide a more nuanced and representative picture of the diaspora as a whole, in comparison to the key informant interviews and focus group discussions that are typically used to capture this type of information. In particular, the data can provide information on the interests and preferences of different segments of the diaspora population – based on, inter alia, geography, gender or age, allowing governments and other stakeholders to develop more focused policies, programmes and communications to support their diaspora engagement strategies.

Additionally, web analytics is a remarkably quick and cost-efficient way of capturing actionable information on the diaspora. In many cases, media outlets may provide the requested data at no cost at all, and all that is required is several hours of the researcher's time to analyse and make sense of the data. While it is not a perfect proxy for population data, it provides sufficient accuracy and information for governments to confidently act and make decisions relating to their diaspora engagement strategies. In fact, in this sense, the data may even be more relevant for decision-making than population or migration data: diaspora engagement strategies typically are not and should not be interested in communicating with *all* the diaspora. Given budgetary restraints and communications challenges, diaspora engagement should initially focus on communicating with diaspora members who express a connection to and interest in their home country. The fact that individuals are viewing websites from their country of origin acts as an indicator of this interest, and thus, this group should be considered a primary target for any diaspora engagement strategy.

However, there remain a number of challenges and limitations with web analytics as a method for diaspora mapping that must be acknowledged and taken into account. First of all, access to the data is not a given and may require significant amounts of liaising and relationship-building with relevant media companies. While in some countries, media outlets may be more open to sharing their data, the culture of information-sharing for research purposes is less developed in other countries. In

Azerbaijan, for example, we were not able to identify a news or media site that was willing to share its web analytics data with our researchers. Second, there are anomalies in the data that need to be understood in order to interpret it accurately. For example, it turns out that despite what the data says, Coffeyville, Kansas, in the United States, does not have a large Armenian population. It just happens to be the default centre starting point for Google Maps in the United States. As such, when a website view comes from a location within the United States that cannot be accurately located, it gets listed as being from the default centre of the United States – Coffeyville, Kansas. Likewise, if there are major events that could result in a sudden increase in tourism in a country or city (Olympic Games or World Cup, for example), these also need to be taken into account when interpreting the data.

Finally, it should also be noted that not all segments of the diaspora population have equal access to the Internet. Particularly, elderly populations may not access the Internet regularly, and this may skew the data towards a younger audience. However, as greater shares of the population are continuously becoming more familiar and comfortable with accessing the Internet and costs for Internet access continue to fall, this last point has become less of an issue in recent years.

While these tools are not without their faults or risks, we have only just begun to scratch the surface of how these and other big data analysis techniques may be used to add insights into migrant and diaspora communities, which can add significant value to diaspora mapping and how governments may better understand, serve and communicate with their diaspora populations. There are promising research opportunities to refine further and improve the work and methodology so that it can be integrated and mainstreamed into diaspora mapping projects.

REFERENCES*

Alinejad, D., L. Candidatu, M. Mevsimler, C. Minchilli, S. Ponzanesi and F.N. van der Vlist
2019 Diaspora and mapping methodologies: Tracing transnational digital connections with "mattering maps". *Global Networks*, 19(1):21–43.

Böhme, M., A. Gröger and T. Stöhr
2020 Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142.

Demilt, J.
n.d. The origins of Twitter. Pennington Creative. Available at https://penningtoncreative.com/the-origins-of-twitter/.

Diaspora Liaison Office of Zambia
2011 *Zambian Diaspora Survey: Report Feeding into the Development of a Diaspora Engagement Framework for Zambia*. Available at www.iom.int/sites/g/files/tmzbdl486/files/jahia/webdav/shared/shared/mainsite/activities/countries/docs/zambia/Zambian-Diaspora-survey-Report.pdf.

Drašković, D.
2020 Can Google search queries help countries map their diasporas? United Nations Development Programme (UNDP), 18 March. Available at www.rs.undp.org/content/serbia/en/home/blog/2020/mo_e-li-Google-pretraga-da-pomogne-u-lociranju-dijaspore-.html.

European Union Global Diaspora Facility (EUDiF)
n.d. Diaspora engagement map. Available at https://diasporafordevelopment.eu/interactive-map/.

Fabos, A.H., L. Kahn and M. Sarkis
2021 Moving stories: Methodological challenges to mapping narratives and networks of people in diasporas, *Journal of Refugee Studies*, 34(3):2554–2567.

Gevorkyan, A.V.
2019 Lessons from an Armenian diaspora online survey: A diaspora portal and non-monetary development initiatives in small economies. *EVN Report*, 16 May. Available at www.evnreport.com/raw-unfiltered/lessons-from-an-armenian-diaspora-online-survey.

** All hyperlinks were working at the time of writing this report.*

iDiaspora, Global Research Forum on Diaspora and Transnationalism (GRFDT), Centre for Research on North America (CISAN) and Africa-Europe Diaspora Development Platform (ADEPT)

2021 *Maximizing Diaspora Engagement: Building Trust, Mobilizing Resources and Ensuring Sustainability*. Insights and reflections paper. IOM, Geneva. Available at https://publications.iom.int/books/virtual-exchanges-maximizing-diaspora-engagement-building-trust-mobilizing-resources-and.

International Organization for Migration (IOM)

2010 *Angola: A Study of the Impact of Remittances from Portugal and South Africa*. Migration Research Series No. 39. Geneva. Available at https://publications.iom.int/books/mrs-no-39-angola-study-impact-remittances-portugal-and-south-africa.

2017 *Migration in the 2030 Agenda*. Geneva. Available at https://publications.iom.int/books/migration-2030-agenda.

2019 *Glossary on Migration*. International Migration Law No. 34. Geneva. Available at https://publications.iom.int/books/international-migration-law-ndeg34-glossary-migration.

2021 *Skills Mapping Through Big Data: A Case Study of Armenian Diaspora in the United States of America and France*. Yerevan. Available at https://publications.iom.int/books/skills-mapping-through-big-data-case-study-armenian-diaspora-united-states-america-and-0.

2022 *Diaspora Mapping Toolkit*. Geneva. Available at https://publications.iom.int/books/diaspora-mapping-toolkit.

IOM and Migration Policy Institute (MPI)

2012 *Developing a Road Map for Engaging Diasporas in Development: A Handbook for Policymakers and Practitioners in Home and Host Countries*. Geneva and Washington, D.C. Available at https://publications.iom.int/books/developing-road-map-engaging-diasporas-development-handbook-policymakers-and-practitioners.

IOM, UNDP and Swiss Agency for Development and Cooperation (SDC)

2021 Diaspora Mapping and Engagement: Global Webinar. IOM–UNDP Global Programme on Making Migration Work for Sustainable Development (Phase III). Synthesis report. Available at https://migration4development.org/sites/default/files/2022-02/Diaspora%20Mapping%20and%20Engagement%20Global%20Webinar.pdf.

Jones, M.V., N. Coviello and Y.K. Tang.

2011 International entrepreneurship research (1989–2009): A domain ontology and thematic analysis. *Journal of Business Venturing*, 26(6):632–659.

Routed Magazine and iDiaspora

2021 *Empowering Global Diasporas in the Digital Era*. Available at www.idiaspora.org/sites/g/files/tmzbdl181/files/resources/document/routedidiaspora2021empowering-global-diasporas-in-the-digital-era.pdf.

Schwartz, R.

2007 Exploring the link between Moldovan communities abroad (MCA) and Moldova. Swedish International Development Cooperation Agency (Sida) and IOM. Available at www.yumpu.com/en/document/view/51297153/exploring-the-link-between-moldovan-communities-abroad-iom.

Smith, T.

2000 *Foreign Attachments: The Power of Ethnic Groups in the Making of American Foreign Policy*. Harvard University Press, Cambridge, Massachusetts.

Taylor, J., J. Rubin, C. Giulietti, C. Giacomantonio, F. Tsang, A. Constant, L. Mbaye, M. Naghsh Nejad, K. Kruithof, M. Pardal, A. Hull and T. Hellgren

2014 Mapping diasporas in the European Union and the United States. IZA Research Report No. 64. RAND Europe. Available at https://ftp.iza.org/report_pdfs/iza_report_64.pdf.

United Nations Economic Commission for Europe (UNECE)

2012 Towards common definition and measurement of diaspora: Practices and lessons from South-Eastern, Eastern Europe and Central Asia. Available at https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2012/WP_25_IOM_Manke_REV.pdf.

Williams, S. and M. Ralphs

2013 Preliminary research into Internet data sources. Government Statistical Services. Available at www.unglobalpulse.org/wp-content/uploads/old_site/williamsralphs.pdf.

ADDITIONAL READING

Kuznetsov, Y.

2008 Mobilizing intellectual capital of diasporas: From first movers to a virtuous cycle. *Journal of Intellectual Capital*, 9(2):264–282.

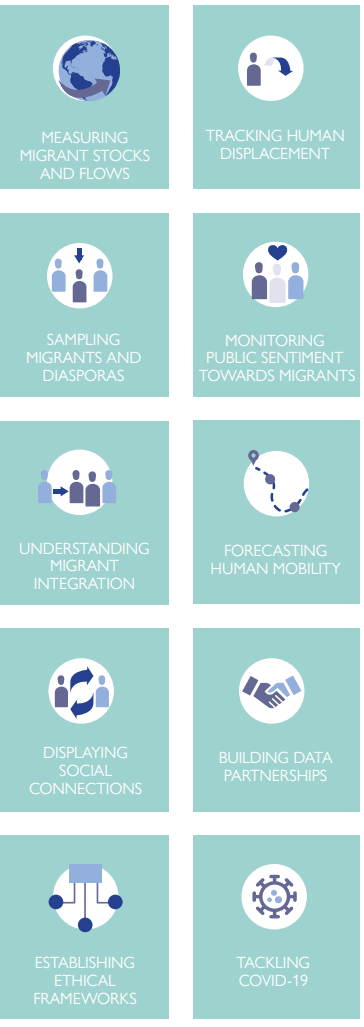
Organisation for Economic Co-operation and Development (OECD)

2015 *Connecting with Emigrants: A Global Profile of Diasporas 2015*. Available at www.oecd-ilibrary.org/social-issues-migration-health/connecting-with-emigrants_9789264239845-en.


71277
26722
93499
37833
04500
82388
59055
60166
15611

CHALLENGES OF HARNESSING DATA INNOVATION FOR MIGRATION POLICY AND RESEARCH

11



THE CHALLENGES OF USING NEW DATA SOURCES AND METHODS FOR MIGRATION ANALYSIS AND POLICY

Niklas Sievers[1] *and Marzia Rango*[2]

Introduction

What are the challenges to face collecting data on human mobility through new technologies and data sources?

ETHICAL CHALLENGES

Earning public trust in the face of new data privacy and security risks

QUALITY CHALLENGES

Distilling accurate information from biased and erroneous data

DATA AVAILABILITY CHALLENGES

Ensuring effective and secure data access across sectors and countries

DATA ANALYTICS CHALLENGES

Integrating requirements of traditional and innovative sources and methodologies

[1] At the time of writing, Niklas Sievers is Data Knowledge Officer at IOM's Global Migration Data Analysis Centre (GMDAC). He developed several research projects exploring the potential and pitfalls of innovative data sources, methods, and tools for migration policy and research. Before joining the United Nations System, he worked as an Advisor in the European Union Advisory Unit of PricewaterhouseCoopers and as a Lecturer at Humboldt University. He holds an MSc from the London School of Economics and Political Science in the social sciences and two bachelor's degrees from Leuphana University.

[2] At the time of writing, Marzia Rango was leading the work on data innovation, capacity-building and analytics at IOM GMDAC in Berlin. She is the co-convenor of the Big Data for Migration Alliance, a joint initiative of GMDAC, the European Commission's Joint Research Centre and the Governance Lab at New York University, seeking to accelerate the responsible use of new data sources and methods for migration analysis and policy. She now works as a Migration and Human Mobility Specialist at the United Nations Operations and Crisis Centre in New York.

The thematic applications outlined in this handbook have demonstrated a wide range of opportunities for new data sources and innovative methods to provide timely information relevant to migration policy. Whether it may be estimating cross-border and internal human mobility, measuring migrants' socioeconomic integration processes, or forecasting migration flows, policymakers and practitioners working on topics related to migration and mobility would benefit from complementary insights gained by leveraging the opportunities offered by new technologies. At the same time, several challenges have been highlighted that require careful consideration. This chapter summarizes the main difficulties and caveats, discussing the challenges along four key categories: ethics and data responsibility, data quality, data availability, and data analytics.

Ethics and data responsibility: Earning public trust in the face of new data privacy and security risks

The ethical risks of using new data sources and innovative methods in migration are among the most widely discussed. Given the novelty of harnessing new technologies to extract data relevant to policy and programming, practitioners across sectors face the need to earn public trust, especially since most of these data are collected by for-profit companies and not always under the full awareness of the data subjects (OSCE, 2021). In addition, various high-profile incidents highlighted how unethical data usage could exacerbate discrimination against certain groups of individuals (Zuiderveen Borgesius, 2018), facilitate targeted misinformation (Gibney, 2018) and support political persecution (UNHCR, 2021). Critical questions have been raised about the effectiveness of existing safeguards against rapid technological advancements, and concerns have been voiced over the use of private data sources and artificial intelligence-based systems for migration policymaking (Beduschi, 2020; Molnar, 2019). While new normative standards – such as the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (European Commission, 2021) – are being developed at least in some parts of the world, practitioners should lead by example, building on existing guidance and establishing good practices to protect all individuals and communities who are or could be represented by data derived from new technologies.

The ethical risks mentioned in this handbook can be organized into three areas: data privacy, bias and discrimination, and data security. First, regarding *data privacy*, the use of new data sources in migration policy without effective safeguards risks revealing private information potentially undermining migrants' fundamental rights (Kuner and Marelli, 2020). A few characteristics inherent to innovative methods exacerbate these risks as they are difficult to align with established data privacy concepts:

(a) *The sheer amount of data collected and processed through new technologies*. Whereas the collection of data from traditional sources follows the principle of data minimization – i.e. to collect only the personal data that are relevant and necessary to reach a specific purpose – new data sources are generated automatically through the use of digital devices. Therefore, they are not collected according to predefined principles (UNSDG, 2017). The "oversupply" of data might lead to further consequences, such as the reidentification of anonymized data (Kuner and Marelli, 2020), including through triangulation of various data sources, and the repurposing of data from harmless to harmful contexts – for instance, information about hospital locations in war contexts (Dodgson et al., 2020).

(b) *The difficulty in obtaining informed consent for using data from new technologies for specific policy and research purposes*. Depending on the modes of data collection, it may not be possible for researchers to contact data subjects to obtain consent at any moment of the data life cycle (Dodgson et al., 2020; UNSDG, 2017). Further, under normative frameworks for traditional data, individuals provide permission to process their data only for the purposes specified when consent was obtained (Kuner and Marelli, 2020). However, the manifold opportunities from the volume, velocity and granularity of data from new technologies may lead to repurposing data sets and conducting analyses that were not defined when consent was given (Kuner and Marelli, 2020; UNSDG, 2017).

(c) *The different incentive structures in the public and private sectors*. Whereas migration policy practitioners and researchers insist on the legibility of algorithms and the need to provide clear explanations of findings to partners and data subjects, in the private sector algorithms are considered proprietary information as they are inextricably linked to revenues and organizational well-being (OCHA, 2020).

Second, automated analytical methods can exacerbate ethical risks since they can amplify and transmit existing bias in the data into real-life decisions, such as in the case of automated job application screenings and border passport control systems (O'Neil, 2016). This has implications in terms of not only discriminatory practices but also individuals' fundamental right of access to justice (see article by Reichel and Molnár), as given the opaqueness of certain algorithms, it may be difficult to challenge the resulting decisions, particularly if no recourse to a human decision maker is provided (Leslie, 2019).

Third, new data methods can exacerbate risks related to *data security*. Typically, the organizations collecting data through new technologies possess the resources necessary to establish secure data infrastructures. However, security precautions may vary, and breaches can occur (OCHA, 2020). Also, technologies used for sharing can heighten the data security risks, particularly when data are accessible on a cloud (UNSDG, 2017). This is particularly relevant in the context of sharing of private data for purposes of research or public policy. The failure to implement effective data security precautions would expose organizations and migrants to severe risks, since unsecured data can harm migrants, particularly if the breached data hold information on the locations and identities of persecuted individuals (OCHA, 2020; Kuner and Marelli, 2020). Data breaches would also imply serious reputational risks to organizations, which can lead to the loss of public trust and undermine the impact of further (not-)for-profit operations.

Data quality: Distilling accurate information from biased and erroneous data

The second group of challenges relate to the quality of data obtained through new data sources and methods. These revolve mainly around bias and representativeness, and errors and uncertainties in the data derived from new technologies.

Regarding challenges of bias and representativeness, the thematic applications have outlined the risk of mistaking the very large sample sizes of certain kinds of big data, such as social media or mobile phone data, for the whole population (see articles by Rowe et al. and the United Nations Global Working

Group on Big Data for Official Statistics). However, despite the relatively large sample size, the data collected through the use of digital devices represent a selected group of individuals, as access to such devices may often vary across gender, age and socioeconomic backgrounds (Mayer-Schönberger and Cukier, 2013). Different social media platforms target different population groups; for instance, Facebook and Instagram are generally more popular among younger demographics, whereas Twitter and LinkedIn are mostly used by working individuals for professional purposes (see article by Bircan et al.). Additionally, the analysis of geographic information from social media data can be challenging (see article by Newson and Sievers). Facebook, for instance, does not list certain countries as countries of origin, and Twitter provides geolocated data only for a minority of all tweets.

Applications based on data from mobile network operators (MNO) have highlighted further data quality challenges. For instance, call detail records (CDRs) or external data representations (XDRs) indicate presence only during communications, meaning when users send or receive messages, calls, or data packages. Hence, individuals rarely using their phones, such as those in older age cohorts, are less represented (see article by Tatem et al.). In addition, registered users may not correspond to actual users; for instance, in some countries, phones used by women tend to be registered in the names of their husbands or sons (see article by Bircan et al.). Despite the widespread use of mobile phones in most countries around the world, the subscriber bases of MNOs may also not be representative of the wider society, especially if MNOs have strong marketing to attract specific target groups or exclude certain groups due to legal reasons – for instance, those under the age of 18. Lastly, the thematic applications highlighted that MNO data are better suited to analyse internal migration and mobility, or patterns of transnationalism; however, users might change SIM cards when travelling or moving abroad, making it harder to use these data to measure international migration (see article by the United Nations GWG on Big Data).

Further, new data sources and methods imply a range of uncertainties and potential errors in the data. The application programming interfaces (APIs) of social media companies like Meta and Twitter can change over time, along with the variables, definitions and algorithms used to categorize "migrants". This creates challenges for longitudinal studies and historical comparisons, which are even more difficult as some social media services allow obtaining estimates only when querying the data (see article by Kim et al.). In addition, faulty data such as fake accounts, double accounts, bots, and spam activities further undermine the quality of data, although some companies provide estimates of their pervasiveness (fake or double accounts on Facebook, for instance, account for approximately 5 per cent of monthly active users globally – see article by Pötzschke). Lastly, mobile phone data might also include noise – for instance, due to unintended roaming in the border area between two countries (see article by the United Nations GWG on Big Data).

Data availability: Ensuring effective and secure data access across sectors and countries

Despite the progress in traditional migration statistics over the past decades, information gaps persist for certain migration-related phenomena, particularly across countries with weaker statistical systems. This is precisely one of the strongest arguments to harness existing (private) data and tools. At the same time, data offered by new technologies are not always freely accessible for use in migration statistics (e.g. MNO data), and practitioners may often find it hard to access data from private entities.

The main challenges in accessing non-traditional data are due to the difficulties in coordination across relevant sectors and organizations (see articles by Pötzschke, and Newson and Sievers). In particular, sharing data between for-profit MNOs and public entities requires multilayered arrangements to ensure a secure, ethical and effective collaboration for both sides. These arrangements imply delicate trade-offs between privacy protection, research and policy purposes. Meanwhile, open data access would offer possibilities to develop detailed indicators and share fully anonymized and aggregated data, guaranteeing the privacy of users more effectively (see article by the United Nations GWG on Big Data). A range of practices and tools can help solve these challenges, such as storing the data on the servers of a trusted data intermediary, establishing good ethical practices and legal frameworks, and collaborative data-sharing agreements (see article by Verhulst and Young). While these have been tested and used in several policy areas, including migration and human mobility, particularly since the beginning of the COVID-19 pandemic (see article by Vespe et al.), more investments and experimentation are needed to scale them up globally, be it at the local, national or cross-country level.

Data analytics: Integrating the requirements of traditional and non-traditional sources and methods

Even when vast volumes of data from new data sources might be available and methods to correct their biases implemented, extreme caution should be exercised in the analysis and interpretation of results. To harness the potential of these sources effectively and correctly, practitioners and researchers will need to adapt concepts, skills and institutional processes in various ways. Applications described in this volume highlighted that existing migration concepts and definitions can be difficult to apply when exploiting new data sources and innovative methods. Practitioners and researchers need to understand how, for instance, social media and mobile phone services define a user's country of residence (see article by Kim et al.). New avenues for combining traditional and non-traditional data sources will need to be explored in a way that is demand-driven and can respond to specific research, programming or policy purposes.

The interoperability between new and traditional data sources and methods is another challenge. Data across new and traditional sources can come in all shapes and sizes, making their integration challenging, especially if the data do not provide geographical information to serve as a framework to overlay and link data sets together (Wesolowski et al., 2014; Ruktanonchai et al., 2018). Further, deriving relevant insights from vast volumes of big data requires a specific set of hard and soft skills. Most data professionals may possess advanced computing and data engineering skills, but contextual domain expertise is just as necessary to make meaningful interpretations and communicate insights and inconsistencies from the use of new data sources for migration policy (Daas et al., 2012:258–259). In this regard, practitioners require budgets and resources for considerable improvements (ibid.). Interdisciplinary approaches will be particularly important to make advancements in this field.

Conclusion

The thematic applications introduced in this handbook have demonstrated a wide range of opportunities for enhancing the use of new data sources, methods, and tools to better understand migration-related phenomena in support of programming and policymaking. This article has highlighted the key challenges in terms of ethics and data responsibility, data quality, data availability,

and data analytics, providing a concise account of the main difficulties and caveats, and pointing out ways to incorporate and potentially address those issues. The use of mobile phone data can provide rich insights for estimating migrant stocks and flows, primarily within the same country, yet its main caveat is the scarcity of demographic data. Social media, on the other hand, offers vast amounts of demographic information, but data can be biased, erroneous and uncertain. Lastly, satellite imagery can valuably support conducting national censuses, especially in remote areas. Yet estimates based on these innovative data cannot provide information about individual intentions and motivations, or patterns of migration journeys of individuals or communities.

The next part of the handbook will put forward several hands-on approaches to support practitioners in developing effective data innovation applications going forward.

REFERENCES*

Beduschi, A.

2020 International migration management in the age of artificial intelligence. *Migration Studies*, 9(3):576–596.

Daas, P.J.H., M.J. Puts, B. Buelens and P.A.M. van den Hurk

2012 Big data as a source for official statistics. *Journal of Official Statistics*, 31(2):249–262.

Dodgson, K., P. Hirani, R. Trigwell and G. Bueermann

2020 *A Framework for the Ethical Use of Advanced Data Science Methods in the Humanitarian Sector*. Data Science and Ethics Group. Available at www.migrationdataportal.org/resource/framework-ethical-use-advanced-data-science-methods-humanitarian-sector.

European Commission

2021 Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM/2021/206. Brussels. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

Gibney, E.

2018 The scant science behind Cambridge Analytica's controversial marketing techniques. 29 March. Available at www.nature.com/articles/d41586-018-03880-4.

Kuner, C. and M. Marelli (eds.)

2020 *Handbook on Data Protection in Humanitarian Action*. Second edition. International Committee of the Red Cross, Geneva. Available at www.icrc.org/en/data-protection-humanitarian-action-handbook.

Leslie, D.

2019 Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. Available at https://doi.org/10.5281/zenodo.3240529.

Mayer-Schönberger, V. and K. Cukier

2013 *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. John Murray, London.

Molnar, P.

2019 Technology on the margins: AI and global migration management from a human rights perspective. *Cambridge International Law Journal*, 8(2):305–330.

** All hyperlinks were working at the time of writing this report.*

Office of the United Nations High Commissioner for Refugees (UNHCR)
2021 News comment: Statement on refugee registration and data collection in Bangladesh. 15 June. Available at www.unhcr.org/news/press/2021/6/60c85a7b4/news-comment-statement-refugee-registration-data-collection-bangladesh.html.

O'Neil, C.
2016 *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, New York.

Organization for Security and Co-operation in Europe (OSCE)
2021 *Principles on Identification for Sustainable Development: Toward the Digital Age*. Available at www.osce.org/odihr/496387.

Ruktanonchai, N.W., C.W. Ruktanonchai, J.R. Floyd and A.J. Tatem
2018 Using Google Location History data to quantify fine-scale human mobility. *International Journal of Health Geographics*, 17:28.

United Nations Office for the Coordination of Humanitarian Affairs (OCHA)
2020 Note #3: Data responsibility in public–private partnerships. 3 February. Available at https://centre.humdata.org/guidance-note-data-responsibility-in-public-private-partnerships/.

United Nations Sustainable Development Group (UNSDG)
2017 *Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda*. Available at https://unsdg.un.org/resources/data-privacy-ethics-and-protection-guidance-note-big-data-achievement-2030-agenda.

Wesolowski, A., G. Stresman, N. Eagle, J. Stevenson, C. Owaga, E. Marube, T. Bousema, C. Drakeley, J. Cox and C.O. Buckee
2014 Quantifying travel behavior for infectious disease research: A comparison of data from surveys and mobile phones. *Scientific Reports*, 4:5678.

Zuiderveen Borgesius, F.
2018 Discrimination, artificial intelligence, and algorithmic decision-making. Council of Europe, Strasbourg. Available at https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.

ADDITIONAL READING

Albertinelli, A., P. Alexandrova, C. Melachrinos and T. Wilkin
2020 Forecasting asylum-related migration to the European Union, and bridging the gap between evidence and policy. *Migration Policy Practice*, X(4):35–41. Available at https://publications.iom.int/books/migration-policy-practice-vol-x-number-4-september-december-2020.

Andreas, P.

2003 Redrawing the line: Borders and security in the twenty-first century. *International Security*, 28(2):78–111.

European Union Agency for Fundamental Rights (FRA)

2018 #BigData: Discrimination in data-supported decision making. 30 May. Available at https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making.

Facebook

2022 Account integrity and authentic identity. Available at www.facebook.com/communitystandards/misrepresentation.

Hathaway, J.C. and M. Foster

2014 *The Law of Refugee Status*. Cambridge University Press.

International Monetary Fund (IMF)

2009 *Balance of Payments and International Investment Position Manual*. Sixth edition. Washington, D.C.

National Academies of Sciences, Engineering, and Medicine (NASEM)

2013 *Frontiers in Massive Data Analysis*. The National Academies Press, Washington, D.C.

2020 *Evaluating Data Types: A Guide for Decision Makers using Data to Understand the Extent and Spread of COVID-19*. The National Academies Press, Washington, D.C.

O'Neil, C. and R. Schutt

2013 *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, Sebastopol.

Pew Research Centre

2017 The digital footprint of Europe's refugees. 8 June. Available at www.pewglobal.org/2017/06/08/digital-footprint-of-europes-refugees.

Pötzschke, S. and B. Weiß

2020 Preliminary findings of the German Emigrants Overseas Online Survey. GESIS – Leibniz Institute for the Social Sciences, Mannheim.

United Nations

2014 The right to privacy in the digital age. A/RES/68/167. Available at https://digitallibrary.un.org/record/764407.

Wesolowski, A., C.O. Buckee, D.K. Pindolia, N. Eagle, D.L. Smith, A.J. Garcia and A.J. Tatem

2013 The use of census migration data to approximate human movement patterns across temporal scales. *PLOS ONE*, 8(1):e52971.

II

DATA INNOVATION GOVERNANCE: ETHICAL FOUNDATIONS AND REGULATORY FRAMEWORKS

ETHICAL FOUNDATIONS OF DATA INNOVATIONS FOR MIGRATION POLICY

12

ARTIFICIAL INTELLIGENCE/ MACHINE LEARNING

ESTABLISHING ETHICAL FRAMEWORKS

BIG DATA, BIG RESPONSIBILITY – FUNDAMENTAL-RIGHTS IMPLICATIONS OF USING ARTIFICIAL INTELLIGENCE IN MIGRATION MANAGEMENT: A EUROPEAN PERSPECTIVE

David Reichel[1] *and Tamás Molnár*[2]

Abstract

This contribution provides general guidance on the fundamental-rights compliant use of big data and artificial intelligence. In view of the increasing uptake of big data analytics and artificial intelligence in the area of migration and border management, many ethical and legal issues arise. This chapter starts by framing the discussion around human rights law. This rights-based approach, anchored in law, is the starting point for any ethical and lawful approach to using big data. Apart from the use of big data for the production of statistics, new technological developments often go one step further by trying to use data to automate certain tasks through predictions based on data and machine-learning algorithms – often referred to and discussed under the broad heading of "artificial intelligence". The contribution outlines main fundamental-rights challenges in relation to the use of artificial intelligence, including data protection and privacy, equality, and non-discrimination, as well as access to justice. This is followed by a brief overview of the potential and actual use of big data and artificial intelligence by border management authorities in the European Union. The article concludes with some practical guidance on using artificial intelligence in a fundamental-rights-compliant manner, drawing on findings from publications by the European Union Agency for Fundamental Rights (FRA).

Introduction: A rights-based approach to big data and artificial intelligence

The collection and availability of data have increased enormously over the past decade. Due to the exponential improvement of computational power, people can use a variety of applications on their smartphones while connected to the Internet. The increased availability of data has sparked research on how such data can be used. Most notably, the

[1] David Reichel works in the Justice, Digital and Migration Unit of the European Union Agency for Fundamental Rights (FRA). He holds a doctorate in Sociology from the University of Vienna, where he taught many courses on migration, human rights and quantitative methodology. Before joining FRA in 2014, David worked at the International Centre for Migration Policy Development.

[2] Tamás Molnár works as Legal Research Officer for the Justice, Digital and Migration Unit of FRA in Vienna. He has taught international law and European Union migration law for 15 years at Corvinus University of Budapest. He studied law in Budapest and Brussels and holds a PhD and habilitation (Dr. habil) in international law (Budapest).

past years have seen considerable advances in the development of machine learning, which usually refers to statistical methods used to make predictions based on data. Such predictions are then used to automate certain tasks – for example, detecting spam emails or human faces on images. As some of the tasks have previously been carried out by humans only, or even go beyond what humans might be able to do, machine learning is considered a strand of artificial intelligence.[3]

However, it did not take long before issues with using machine learning for certain tasks appeared. For example, an automated chatbot expressed racist sentiment within hours (Johnston, 2017), recruitment algorithms generally preferred men over women (Dastin, 2018), and machine translations showed strong gender bias (Prates et al., 2020).

Such examples led policymakers to raise ethical and legal questions with respect to the use of big data and artificial intelligence. Particularly in Europe, the European Parliament adopted resolutions that highlighted the ethical and fundamental-rights implications (see, for instance, European Parliament, 2017). Following the suggestion from the European Council (2017), the European Commission became active and adopted several initiatives on artificial intelligence in April 2018. These include, among others, a communication on *artificial intelligence for Europe* (European Commission, 2018) and the creation of a High-Level Expert Group on Artificial Intelligence.[4] Since then, there have been major efforts in the European Union to tackle the trustworthy use of artificial intelligence, including a concrete proposal for regulating it.[5] The European Commission proposed a law regulating artificial intelligence in April 2021, which includes a risk-based approach. This means that some instances of artificial intelligence would be forbidden, and others, considered as high-risk, are subject to conformity assessments and need to follow certain extra requirements in relation to data governance, documentation and record-keeping, transparency and provision of information to users, human oversight, robustness, accuracy and security. The list of high-risk use of artificial intelligence is subject to regular assessment, and it includes the category of migration, asylum and border management.

In Europe, a strong, multilayered and well-elaborated legal framework for the protection of fundamental and human rights exists. The international human rights framework has advanced considerably in the Old Continent and beyond in the aftermath of the Second World War. Starting with the Universal Declaration of Human Rights (UDHR) in 1948, the United Nations has developed a large body of treaties to protect people's human rights globally. In addition to United Nations agencies (such as the International Labour Organization (ILO)), regional organizations have equally elaborated a body of legal instruments to safeguard human rights. Most notably, the Council of Europe, with its 47 member States, has adopted the European Convention on Human Rights (ECHR), as the most important pan-European instrument to protect human rights with its own regional court – the European Court of Human Rights (ECtHR). Furthermore, at the level of the European Union, with its 27 member States, the ECHR was the basis for the development of the Charter of Fundamental Rights of the European Union. The Charter applies the same protection as the ECHR does, and it includes further rights, going beyond those of the ECHR. It applies to all European Union institutions in all of their actions and the member States whenever the latter act within the scope of European

[3] For an easy introduction to artificial intelligence and machine learning, see: Onuoha and Nucera, 2018; Samoili et al., 2020.

[4] More information is available at https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai.

[5] More information is available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

Union law (Article 51(1) of the Charter).[6] Fundamental rights are further protected and safeguarded in European Union secondary legislation, such as with data protection law and non-discrimination legislation. In addition, they are included and echoed in national constitutions and other pieces of domestic law.

In light of the foregoing, this multilayered European fundamental-rights framework is essentially the starting point for any debate on the ethical use of artificial intelligence – and more. This contribution provides initial reflections and guidance on selected horizontal fundamental-rights issues linked to the use of big data and artificial intelligence in the area of migration and border management, but also beyond. It draws on research carried out by the FRA.

Challenges to fundamental rights

Knowledge about fundamental rights is often limited in the general population, which is also true for those developing and working with big data and artificial intelligence (FRA, 2020a:58–60). However, fundamental rights provide the best starting point for approaching the use of new data sources and new technologies, as they already outline a commonly agreed-upon and democratically established list of rights and principles. This section outlines selected fundamental rights impacted by the use of artificial intelligence to inform practitioners intending to use big data and artificial intelligence. The selected areas are central to the ongoing discussions and horizontally applicable across a wide array of artificial intelligence use cases from a fundamental-rights perspective. The final section outlines how these rights can be observed and upheld from a practical point of view.

The rights to privacy and data protection

The rights to privacy and data protection are included in Articles 7 and 8 of the Charter and Article 8 of the ECHR. They are further protected by European Union law, most notably the General Data Protection Regulation (GDPR),[7] and at the level of the Council of Europe, the Convention 108 +.[8]

Big data and artificial intelligence are used to make decisions that impact people. This is often based on the use of personal data, meaning "any information relating to an identified or identifiable natural person" (Article 4(1) of GDPR). Importantly, information in data sets might also indirectly identify people, meaning that even if direct identifiers are excluded from a data set – such as the name, address or social security number – people might still be identified. This happens either because the information contained is still very unique to certain people in the data set, or when using the anonymized information in combination with other information. This means that the line between personal and non-personal data may often be blurred. Yet as soon as personal data are used, European data protection law applies.

European data protection law requires users of personal data to follow certain principles and requirements. It requires data users to process data (a) lawfully, fairly and in a transparent manner; (b) with a specific, explicit and legitimate purpose; and (c) in compliance with requirements of data minimization, data accuracy, storage limitation, data security and accountability.

[6] For an overview of the human rights framework from a European perspective, see: Wouters et al., 2019:676–742; FRA, 2012.

[7] More information is available at https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[8] More information is available at www.coe.int/en/web/data-protection/convention108-and-protocol. For an overview of European data protection law, see: FRA, 2018c.

In addition, European data protection law limits the use of personal data for automated individual decision-making, including profiling. This means that decisions that have a legal or otherwise significant impact on people (such as decisions on visa or asylum applications) cannot be fully automated based on data processing, including machine learning or other statistical methods. There are, however, exceptions for situations: (a) where the automated processing is needed to enter into a contract; (b) if the data subject gave explicit consent; or (c) if it is authorized by law and the data subject's rights, freedoms and legitimate interests are appropriately safeguarded (Article 22 of the GDPR). What is more, whenever automated decision-making is used, people have the right to receive meaningful information about the logic involved.

A detailed description of data protection issues, and their link to big data and artificial intelligence, is beyond the scope of this chapter.[9] However, data protection lies at the core of fundamental-rights concerns when using new data sources and artificial intelligence. This is because of the increased amount and granularity of data used, which might easily lead to the identification of individuals, and also because the data can more easily contain sensitive personal information. For example, information on ethnic origin or political opinions is often included in social media data, or biometric data, which can be derived from images of people through facial recognition technology. These categories of data ("sensitive data") are especially protected under European Union data protection law (Article 9(1) of the GDPR).

Whenever personal data are used for migration management purposes, these processes have to strictly follow the requirements stemming from European data protection law.

Equality and non-discrimination

Data from social media and from other data sources often include information linked to protected attributes, such as gender, ethnic origin or sexual orientation. As a consequence, data-supported decision-making can lead to unlawful discrimination and exacerbate existing societal inequalities (FRA, 2018d and 2019b). Article 21(1) of the Charter prohibits discrimination, be it direct or indirect, based on the following grounds: sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation.

People might think that data or machines are neutral with respect to certain attributes. This is not the case, and data-supported decision-making (be it fully automated or not) can lead to unfair and discriminatory treatment, for several reasons. First, data used to inform decision-making might not be representative of certain population groups. For example, research found that gender identification through facial recognition technology was highly accurate for white men, but not at all for Black women (Buolamwini and Gebru, 2018). The main reason for this difference was that the algorithms were mainly developed based on white male faces. When the so-called training data are not representative of all population groups, algorithms might lead to discriminatory outcomes.

[9] See, for example: ICO, 2017; Sartor, 2020.

Second, data might reflect historical bias and discriminatory behaviour. This simply means that if people behave in a discriminatory way, and this behaviour is registered in data collection, any statistics and algorithms based on these data will encode such a behaviour and potentially exacerbate it. This was shown, for example, in the area of predictive policing in the United States of America (Richardson et al., 2019).

Third, discrimination might also happen when certain information is not included in the data, or because protected characteristics are highly correlated with certain outcomes, which are to be predicted. For example, if car accidents are analysed, the data will show significant differences by gender, because men more often (but not always) show riskier driving compared to women. Gender becomes a proxy variable for risky driving behaviour, in the absence of information on actual risky driving behaviour. This type of discrimination needs to be analysed in context.

Importantly, discrimination can also happen if information on protected attributes is not included in a data set. Proxy information – e.g. the colour of a car for gender, or citizenship for ethnic origin – can indicate protected attributes and still lead to discrimination. Generally, users of artificial intelligence do not regularly investigate in detail if their tools may discriminate (FRA, 2020a:68–73), which means that more awareness about this area is needed among users of artificial intelligence.

Access to justice

Finally, people in the European Union have the right to a fair trial and an effective judicial remedy (Article 47 of the Charter). This means that people have the right to challenge before a court any decision made about them. Challenging decisions, which are made with the support of data and artificial intelligence, can be tricky due to the complexity and potential opaqueness of the tools and data used. In order to guarantee access to an effective remedy to people seeking redress, who were harmed through the use of artificial intelligence, the following points need to be duly considered (FRA, 2020a:76):

(a) People need to know that artificial intelligence has been used for a decision taken about them.
(b) People need to know where and how to complain regarding the decision taken.
(c) Users of artificial intelligence need to be able to explain how they arrived at a certain decision (to avoid the "black box" effect).

The last point is necessary because decisions cannot be challenged if they cannot be explained. While complex methods for decision-making might be hard to understand, there is always a way to provide meaningful information about the logic involved – to borrow the language used in the GDPR. Therefore, being able to explain decisions not only stems from providing access to justice to those who are challenging decisions, but is also required to comply with the right to good administration, which is a general principle enshrined in European Union law (FRA, 2020a). It is often argued that certain complex artificial intelligence solutions cannot be explained ("black box" effect). While it is true that some applications are difficult to understand, there is nothing that prevents developers and users of artificial intelligence to provide information about the design of the system, and there are ways to even look into the black box. Information needed to assess an artificial intelligence tool and contribute to its explainability is provided in the final section of this contribution.

Other rights areas

Apart from the three horizontal and interlinked rights areas – data protection, non-discrimination and access to justice – several other fundamental rights are impacted by the use of artificial intelligence. The extent to which fundamental rights are impacted depends on the area in which artificial intelligence is used and the context of its application.

For example, in the area of migration management, using artificial intelligence to support decisions on visa applications or border checks can affect the right to asylum. If (live) facial recognition technology is used in public places, the right to freedom of expression or freedom of assembly might be impacted (FRA, 2019a:29–30). This means that any use of artificial intelligence, in whatever context, should consider the full range of fundamental rights before applying the technology. This might often be a quick exercise, as there are no major issues for some use cases in certain areas, and benefits may sometimes exist (e.g. more transparent and consistent decision-making). For example, public administration is using artificial intelligence tools to adjust scans of applications for social benefits or using speech-to-text tools for automated transcription. Yet for applications involving decisions about people that can have a legal or otherwise significant effect, there may be serious challenges.

Increasing automation in the area of migration and border management and fundamental rights

When zooming in on migration, following Beduschi's insightful classification (2021:1,3), artificial intelligence-driven technology can affect the management of international migration in three main ways: (a) modernizing States' and international organizations' (e.g. the European Union, the Office of the United Nations High Commissioner for Refugees) traditional practices; (b) deepening the existing asymmetries on the international plane between States, in terms of access to cutting-edge technologies and other modern infrastructure; and (c) reinforcing contemporary calls for more evidence-based migration management. The list below primarily focuses on the first and third aspects, also bringing in a couple of examples from these perspectives.

If carefully conceived, implemented and monitored, artificial intelligence could bring substantial opportunities to improve the efficiency of migration management, while safeguarding and eventually strengthening fundamental rights (FRA, 2021:167). For example, artificial intelligence-driven tools could:

(a) Speed up the process for applying for and granting international protection, improving transparency and consistency in decision-making;
(b) Protect from identity fraud and unveil related abuses;
(c) Support authorities in quickly identifying missing or abducted children;
(d) Support policymakers in forming more accurate predictions and developing adequate responses to humanitarian emergencies;
(e) Improve authorities' ability to quickly identify vulnerabilities and protection needs;
(f) Progressively reduce the costs for border and migration management.

As highlighted in the previous section, using artificial intelligence systems and big data implicates a wide spectrum of fundamental rights, which include, but also go beyond, privacy and data protection, non-discrimination and access to remedies – also in the context of migration (Beduschi, 2018:1009–1016; FRA, 2020a:7,57; FRA, 2021:167–168).

"Experiments with new technologies in migration management are increasing" (Molnar, 2020:769), and the European Union is no exception. To further explore the use of artificial intelligence-driven technologies in the areas of European Union border and asylum policies, the European Commission conducted two studies in 2020. One explores the feasibility of developing a common European tool for migration prediction based on artificial intelligence, with the aim to further evidence-based migration management (European Commission, 2020a). Another one analyses the pros and cons of using artificial intelligence in the fields of migration, asylum, borders and security, looking into nine concrete migration-related areas, such as border checks at external borders, issuing short-stay visas and residence permits, as well as granting international protection (for more information and to read on the other areas, see: European Commission, 2020b). Likewise, the European Union agency in charge of the operational management of large-scale European Union information technology systems (eu-LISA) plans to establish a Working Group on Artificial Intelligence to enquire into opportunities for the implementation of artificial intelligence in the justice and home affairs area, specifically within its own mandate (Garkov, 2020). The identified use cases by eu-LISA for future implementation of artificial intelligence-related solutions include enabling automatic screening of visa-free travellers on the basis of predefined risk indicators and screening rules (see the example with European Travel Information and Authorization System (ETIAS) below); increasing the analytical toolset for the huge amount of data collected via interoperability of information technology systems (see the example with the Central Repository for Reporting and Statistics (CRRS) below); and improving and enhancing the accuracy of biometric matching algorithms.

However, the future has already begun. The use of algorithms is already regulated and will be progressively implemented in the upgraded and new generation of large-scale European Union information technology systems and their interoperability. Algorithms will help authorities perform certain tasks, such as improving identity checks using facial recognition (FRA, 2019a:13) and fingerprint matching (for an overview, see: FRA, 2020b:58–59). Algorithms will also support the use of the data stored in these information technology systems to predict risks and produce analysis to assist in decision-making (FRA, 2018b:43–44). The legal instruments setting up these European Union databases and regulating their interoperability explicitly regulate the use of artificial intelligence and envisage specific fundamental-rights safeguards.

Let us briefly mention here two examples. One example is the *screening rules of the ETIAS*, which involve an algorithm that compares the data submitted in a visa-free traveller online application with specific risk indicators corresponding to security, irregular migration or public health risks.[10] Another example is the *CRRS* under the Interoperability Regulations,[11] which is a repository of clearly defined anonymized data relating to individuals whose personal data are stored in European Union

[10] More information is available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018R1240.

[11] More information is available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R0817 and https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R0818.

information technology systems, to obtain customizable reports and cross-system statistics and data for policy, operational, and data quality purposes.[12]

These existing and planned features may impact fundamental rights in various ways, and safeguards need to be duly implemented. For example, the ETIAS screening rules include safeguards to prevent discrimination on protected grounds,[13] and national authorities will ultimately decide whether or not to grant an authorization to travel, thus ensuring human control. Nonetheless, risks of indirect discrimination and unjustified profiling could arise from the combination of criteria: a certain ethnic group typically working in an economic sector (e.g. in agriculture) and from a certain country or city could be refused the authorization to travel (FRA, 2018a:112). Similarly, while the CRRS stores anonymized data, it should not lead to reports suggesting operational actions, which would result in discrimination of certain categories of persons (FRA, 2018b). Challenging those decisions might be difficult, as the person concerned should be able to know to what extent the algorithms influenced a decision affecting her or him, and the procedure to follow in such cases.

Ways forward: Fundamental-rights compliant use of artificial intelligence

Following the suggestions of the FRA report on artificial intelligence and fundamental rights (FRA, 2020a), the following three issues need to be taken into account at a very minimum to ensure a fundamental-rights compliant use of artificial intelligence:

(a) Data need to be processed legally, fully in line with data protection laws and principles.
(b) The use of artificial intelligence must not lead to unfair discrimination, based on any of the grounds listed in the Charter.
(c) People need to be able to challenge decisions made about them.

Depending on the area of application, further fundamental-rights consideration applies. These aspects can be assessed prior to the use of an artificial intelligence (supported) system through fundamental-rights impact assessments. A starting point for proper impact assessments should be transparency about key issues included in the system:

(a) A description of the purpose and context of the system and the legal basis.
(b) A description of the possible harm of using the system.[14]
(c) A description of the technology used, including information on the data used for building the system.
(d) Information on training data for artificial intelligence systems should be provided, which allows assessing potential errors linked to the representation of the population covered, as well as potential measurement errors. As described in FRA's paper on data quality (FRA, 2019b:14–15), the following information is useful for such an assessment:

[12] More information is available at https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX%3A32019R0817.

[13] This includes the person's sex, age, colour, race, ethnic or social origin, genetic features, language, political or any other opinion, religion or philosophical belief, trade union membership, membership of a national minority, property, birth, disability or sexual orientation (ETIAS Regulation Article 33 (5)).

[14] See, for example, FRA's list on possible harms, available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence-annex-2_en.pdf.

(i) Where do the data come from, and who was/is responsible for the collection of the data?
(ii) What information is included in the data set?
(iii) Who is covered in the data, and who is potentially underrepresented?
(iv) Is any information missing in the data set?
(v) What are the time frame and geographical coverage of the data collection?

(e) A description of the accuracy of the system in terms of outcomes based on training data and possible tests and experiments in real-life situations. This should include different rates of accuracy for different population and demographic groups.
(f) A description of compliance with existing standards and potential certifications obtained.

Several tools can support the ethical and legal use of artificial intelligence in line with fundamental rights. For example, the European Commission's High-Level Expert Group on Artificial Intelligence developed a checklist that can be used and adapted for certain purposes.[15]

The use of big data and artificial intelligence is quickly expanding. Given these fast developments, it is important to consider the full scope of fundamental rights before a system is put in place. This will make the use of big data and artificial intelligence in various contexts, including migration analysis and policy, much more trustworthy and of higher quality, which will eventually contribute to the increased uptake. Most notably, looking into potential fundamental-rights issues in data innovation helps better understand what the data and artificial intelligence predictions are about. This would pave the way for a greater understanding of important societal issues, such as migration, and how data innovation can improve the lives of individuals.

[15] More information is available at https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

REFERENCES*

Beduschi, A.

2018 The big data of international migration: Opportunities and challenges for States under international human rights law. *Georgetown Journal of International Law*, 49:981–1017.

2021 International migration management in the age of artificial intelligence. *Migration Studies*, 9(3):576–596.

Buolamwini, J. and T. Gebru

2018 Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research (PMLR)*, 81:1–15.

Dastin, J.

2018 Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 11 October. Available at www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

European Commission

2018 Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe. COM(2018) 237 final, Brussels, 25 April.

2020a *Management Plan 2020*. DG Migration and Home Affairs. European Commission, Brussels. Available at https://ec.europa.eu/info/system/files/management-plan-home-2020_en.pdf.

2020b *Opportunities and Challenges for the Use of Artificial Intelligence in Border Control, Migration and Security*. Volume 1: Main report. Available at https://op.europa.eu/en/publication-detail/-/publication/c8823cd1-a152-11ea-9d2d-01aa75ed71a1/language-en.

European Council

2017 European Council meeting (19 October 2017) – Conclusions, EUCO 14/17, Brussels, 19 October.

European Parliament

2017 European Parliament resolution of 14 March 2017 on fundamental rights implications of big data: Privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI)).

European Union Agency for Fundamental Rights (FRA)

2012 *Bringing Rights to Life: The Fundamental Rights Landscape of the European Union*. Luxembourg. Available at https://fra.europa.eu/en/publication/2012/bringing-rights-life-fundamental-rights-landscape-european-union.

2018a *Preventing Unlawful Profiling Today and in the Future: A Guide*. Luxembourg. Available at https://fra.europa.eu/en/publication/2018/preventing-unlawful-profiling-today-and-future-guide.

** All hyperlinks were working at the time of writing this report.*

2018b *Interoperability and Fundamental Rights Implications: Opinion of the European Union Agency for Fundamental Rights*. FRA Opinion – 1/2018 [Interoperability], Vienna, 11 April. Available at https://fra.europa.eu/en/publication/2018/interoperability-and-fundamental-rights-implications.

2018c *Handbook on European Data Protection Law*. 2018 edition. Luxembourg. Available at https://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law-2018-edition.

2018d *#BigData: Discrimination in Data-supported Decision Making*. Luxembourg. Available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-focus-big-data_en.pdf.

2019a *Facial Recognition Technology: Fundamental Rights Considerations in the Context of Law Enforcement*. Luxembourg. Available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-facial-recognition-technology-focus-paper-1_en.pdf.

2019b *Data Quality and Artificial Intelligence – Mitigating Bias and Error to Protect Fundamental Rights*. Luxembourg. Available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf.

2020a *Getting the Future Right: Artificial Intelligence and Fundamental Rights*. Luxembourg. Available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence_en.pdf.

2020b *Handbook on European Law Relating to Asylum, Borders and Immigration*. Edition 2020. Luxembourg. Available https://fra.europa.eu/en/publication/2020/handbook-european-law-relating-asylum-borders-and-immigration-edition-2020.

2021 *Fundamental Rights Report 2021*. Luxembourg. Available https://fra.europa.eu/en/publication/2021/fundamental-rights-report-2021.

Garkov, K.

2020 The digital transformation of internal security in the EU, AI and the role of eu-Lisa. European Union Agency for the Operational Management of Large-Scale IT Systems in the Area of Freedom, Security and Justice, 15 December. Available at www.eulisa.europa.eu/Newsroom/News/Pages/The-digital-transformation-of-internal-security-in-the-EU-AI-and-the-role-of-eu-LISA.aspx.

Information Commissioner's Office (ICO)

2017 *Big Data, Artificial Intelligence, Machine Learning and Data Protection*. Data Protection Act and General Data Protection Regulation. Available at https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

Johnston, I.

2017 AI robots learning racism, sexism and other prejudices from humans, study finds. *Independent*, 14 April. Available at www.independent.co.uk/life-style/gadgets-and-tech/news/ai-robots-artificial-intelligence-racism-sexism-prejudice-bias-language-learn-humans-a7683161.html.

Molnar, P.

2020 AI and migration management. In: *The Oxford Handbook of Ethics of AI* (Dubber, M.D., F. Pasquale and S. Das, eds.). Oxford University Press, Oxford, pp. 769–787.

Onuoha, M. and D. Nucera

2018 *A People's Guide to AI*. Available at https://alliedmedia.org/wp-content/uploads/2020/09/peoples-guide-ai.pdf.

Prates, M.O.R., P.H.C. Avelar and L. Lamb

2020 Assessing gender bias in machine translation – A case study with Google Translate. *Neural Computing and Applications*, 32(1).

Richardson, R., J.M. Schultz and K. Crawford

2019 Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*, 94:15–55.

Samoili, S., M. López Cobo, E. Gómez, G. De Prato, F. Martínez-Plumed and B. Delipetrev

2020 *AI Watch: Defining Artificial Intelligence – Towards an Operational Definition and Taxonomy of Artificial Intelligence*. Luxembourg. Available at https://publications.jrc.ec.europa.eu/repository/bitstream/JRC118163/jrc118163_ai_watch._defining_artificial_intelligence_1.pdf.

Sartor, G.

2020 The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. European Union, Brussels. Available at www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf.

Wouters, J., C. Ryngaert, T. Ruys and G. De Baere

2019 *International Law: A European Perspective*. Hart Publishing, Oxford.

ADDITIONAL READING

Amnesty International

2021 Ban dangerous facial recognition technology that amplifies racist policing. 26 January. Available at www.amnesty.org/en/latest/press-release/2021/01/ban-dangerous-facial-recognition-technology-that-amplifies-racist-policing/.

European Commission

2020 White paper: On artificial intelligence – A European approach to excellence and trust. COM(2020) 65 final, Brussels, 19 February. Available at https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Liberty

n.d. Resist facial recognition. Available at www.libertyhumanrights.org.uk/campaign/resist-facial-recognition/.

SHARE Foundation, Hermes Center, Bits of Freedom, ARTICLE19, Homo Digitalis and EDRi

2020 Campaign "Reclaim Your Face" calls for a ban on biometric mass surveillance. 12 November. Available at https://edri.org/our-work/campaign-reclaim-your-face-calls-for-a-ban-on-biometric-mass-surveillance/.

13

MOBILE PHONES

SOCIAL MEDIA

ARTIFICIAL INTELLIGENCE/ MACHINE LEARNING

SATELLITE DATA

OTHER

ESTABLISHING ETHICAL FRAMEWORKS

DATA INNOVATION GOVERNANCE IN HUMANITARIAN CONTEXTS: ETHICAL AND REGULATORY FRAMEWORKS

Robert Trigwell,[1] *David Eduardo Zambrano*[2] *and Gretchen Bueermann*[3]

Introduction

While the question of ethical innovation and data use opens a myriad of avenues for conceptual discussions, the central factor to better harness new data sources and analytical methods is inextricably linked to processes and practical considerations while ensuring the protection of displaced and mobile persons. This chapter guides readers through the concepts developed by the Humanitarian Data Science and Ethics Group (DSEG) and contained in *A Framework for the Ethical Use of Advanced Data Science Methods in the Humanitarian Sector* (Dodgson et al., 2020). In particular, the chapter explains the step-by-step process and links between the application of the project and the principles proposed in the Framework using a case study from IOM operations in Nigeria to support the region's recovery planning. This case study illustrates how safe and responsible data innovation practices are central throughout the entire process. New and innovative data collection and analysis methods are hinged on ensuring adequate data protection, assessing bias in different models, documenting the processes of those models, making analysis available for peer review, and guaranteeing the greatest possible legibility.

This work contributes to and overlaps with nascent debates on responsible and ethical data collection and use, data protection, humanitarian innovation, and humanitarian principles and standards. What does real informed consent look like in a humanitarian context? Who should have access to humanitarian data? What level of automation is acceptable in decision-making? Like previous technological innovations, there is an increased interest in using more advanced methods of data science in humanitarian work. However, this enthusiasm is matched by a parallel urgency to highlight the incumbent ethical considerations in a practical and accessible way, to support humanitarians to

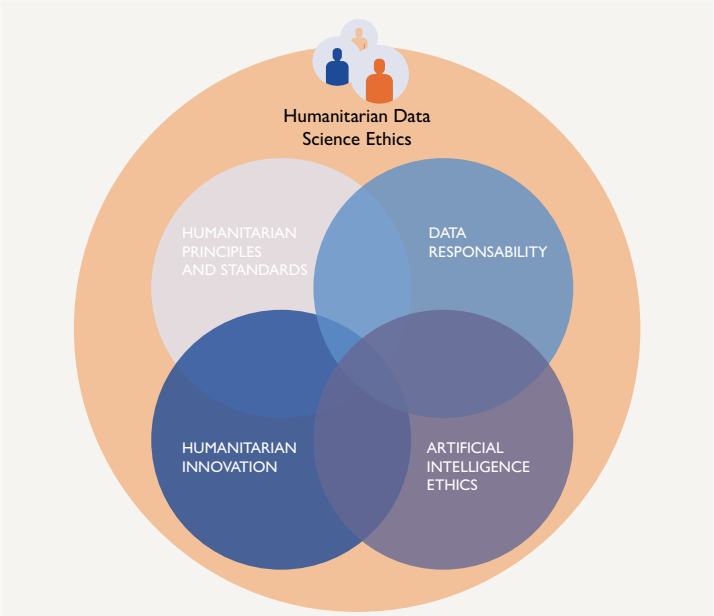
[1] Robert Trigwell has a humanitarian relief background, having worked with non-governmental organizations and the United Nations across the Middle East, East Africa and Asia. Rob currently works within the Global Displacement Tracking Matrix team as Senior Coordination Officer in Geneva.

[2] David Eduardo Zambrano is a former data scientist for IOM's Displacement Tracking Matrix in Geneva, Switzerland. Previously, he was a guest scientist at the Max Planck Institute for the Physics of Complex Systems. Eduardo holds a PhD from the Brazilian Center for Research in Physics.

[3] Gretchen Bueermann is Research and Analysis Specialist at the World Economic Forum's Centre for Cybersecurity. Previously she was a Data Scientist at IOM. Prior to that, Gretchen worked at the Behavioral Research Lab at Yale University. She has a master's in Development Economics from Yale University and a bachelor's degree in Economics from Reed College.

better understand the challenges of introducing innovative technology or data science methods in the sector (McDonald, 2016; Raymond and Al-Achkar, 2016). Although the application of data science projects (especially artificial intelligence projects) to humanitarian operations remains largely in the development or early integration stages, there is an opportunity to proactively and critically assess new technologies introduced to the sector and the resulting implications.

Figure 1. Overlapping areas of humanitarian data science and ethics, as identified by the Humanitarian Data Science and Ethics Group



DSEG[4] was created in 2018 to ensure that any innovative application of data science to support humanitarian outcomes is responsible and safe, and that ethical frameworks inform the process. DSEG brought together a diverse group of humanitarian practitioners, data ethicists, academics and innovators to develop *A Framework for the Ethical Use of Advanced Data Science Methods in the Humanitarian Sector*, which helps practitioners ensure that their projects are executed ethically and responsibly. The group's work is at the juncture of humanitarian principles and standards, data responsibility, artificial intelligence ethics and humanitarian innovation, as demonstrated in Figure 1. By convening a diverse group of stakeholders on this topic, the goal is to bridge the gap between theory and practice.

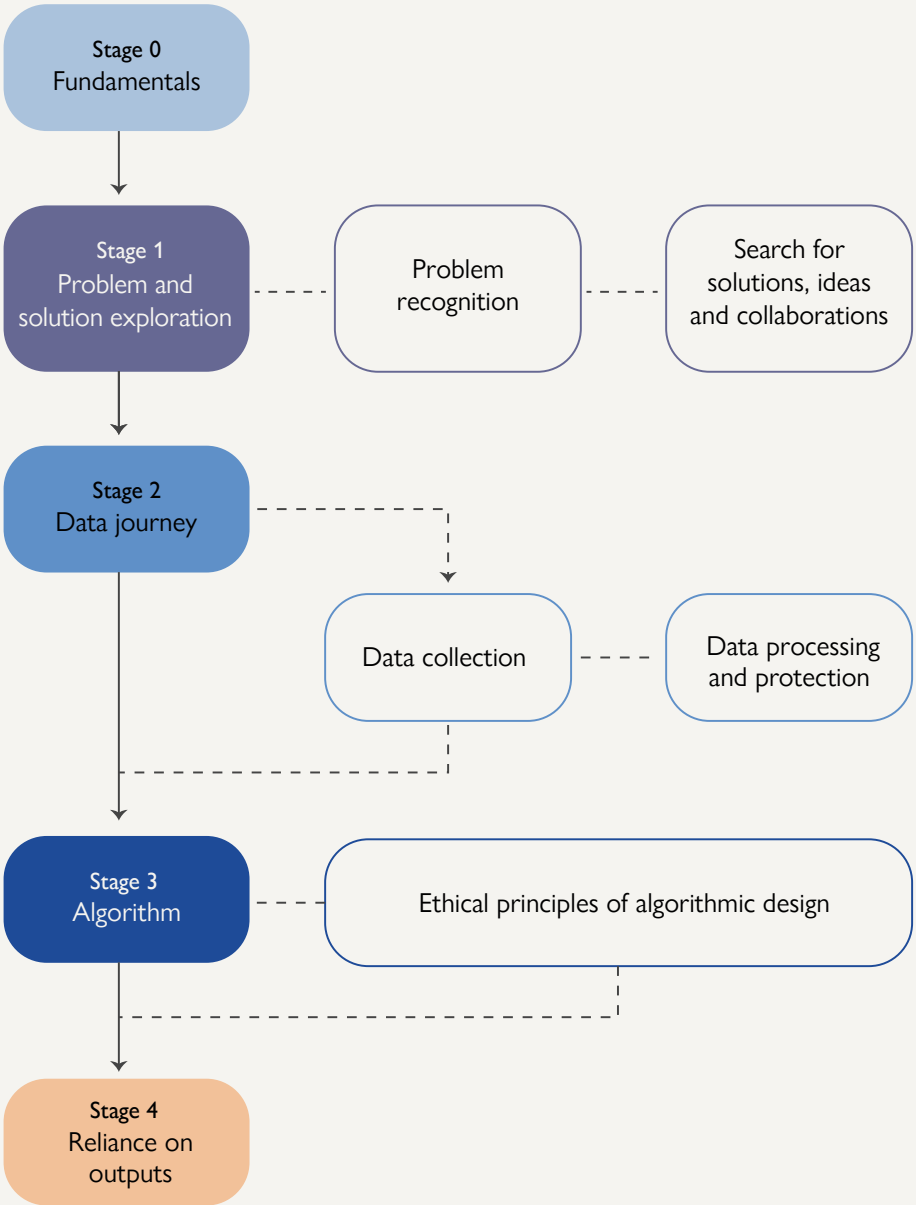
As demonstrated in Figure 2, the Framework covers the fundamental ethics that inform responsible data science use and the five key stages of the data science life cycle:

(0) Fundamentals (humanitarian principles, artificial intelligence ethics, human rights and data management)
(1) Problem and solution recognition
(2) Data checks
(3) Algorithm checks
(4) Output development

[4] More information is available at www.hum-dseg.org/.

Fundamentals are listed as Stage 0, as these are the necessary foundations to build the project on, rather than a step within the process. Problem and solution recognition are part of Stage 1, as these are critical in *any* programme, not limited to one on data science. For any programme, having a full understanding of the problems and challenges at hand and exploring suitable solutions is foundational. This step is designed to mitigate "techno-solutionism" and rather to understand *how* the problem can or cannot be solved using technology. This step aims to avoid data science falling into the bracket of innovating for the sake of innovating. Once the programme team has a full understanding of the problem and has identified the solution to be facilitated or improved with data science implementation, then the following steps become more technical in terms of data science. The Framework is designed to ensure everything is (a) ethically sound and principled, (b) pragmatic and (c) methodologically robust. This way, we can ensure that the data science stakeholders are adhering to what makes the humanitarian sector what it is – the principled-manner approach to their work.

Figure 2. The data science life cycle, as identified by the Humanitarian Data Science and Ethics Group



To highlight why and how this work is important, this chapter will demonstrate how one of IOM's Displacement Tracking Matrix (DTM)'s data science projects was developed in Nigeria using the Ethical Framework as a guide, showcasing an operational example of how a service-delivery-informing model was developed following the steps of the DSEG Ethical Framework. The goal is to demonstrate that the responsible application of data science to humanitarian projects is both feasible and easy to accomplish, when all risks are considered and mitigated, and any limitations of the process are clearly documented as part of the project output.

Returns and stability in the Lake Chad Basin

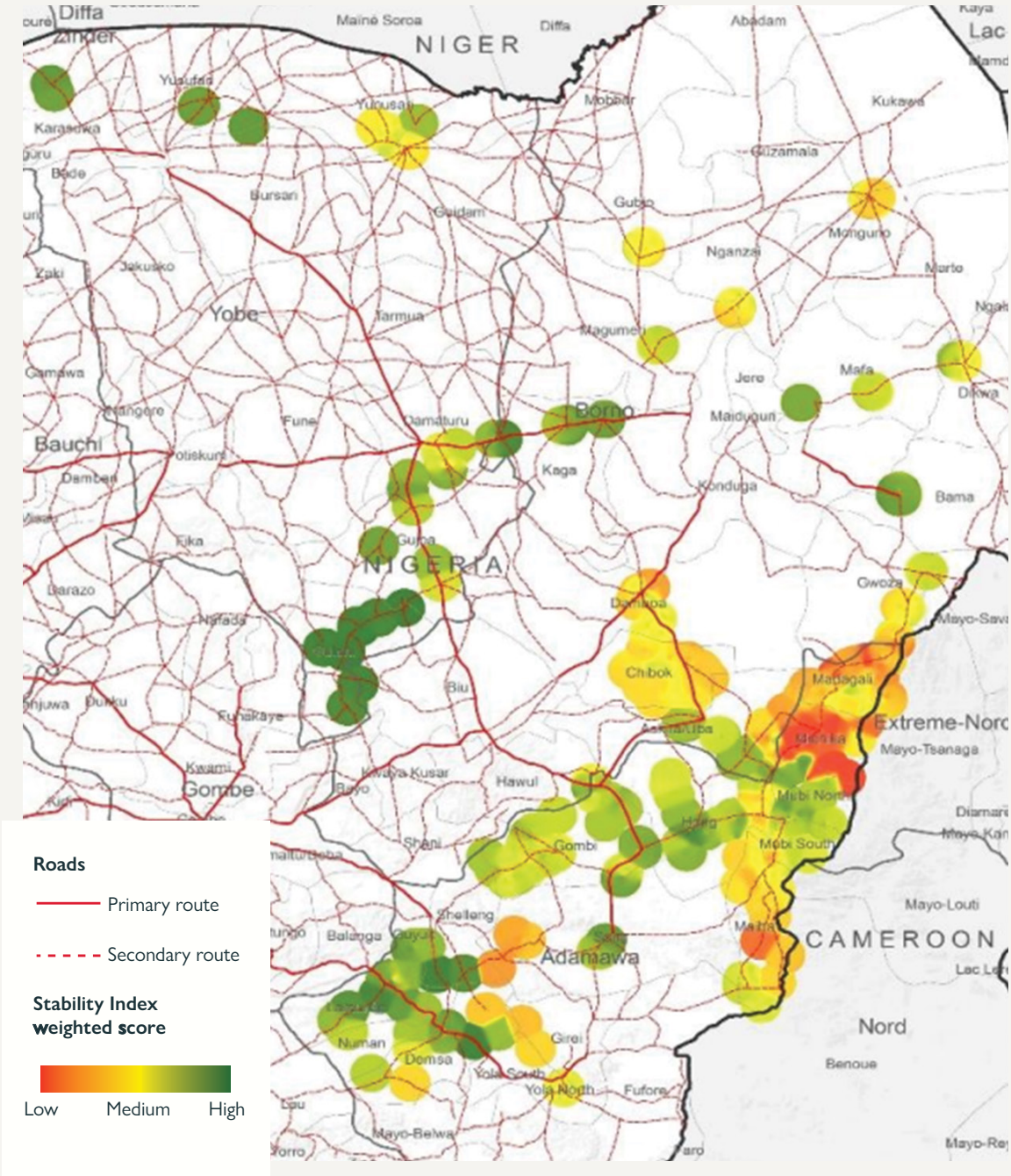
Since 2014, the Lake Chad Basin crisis has affected some of the world's most vulnerable people in Nigeria, Cameroon, the Niger and Chad. The crisis – caused by a complex combination of non-State armed groups, the onset of violent communal clashes and climate change – has led to the forced displacement of nearly 4.5 million people, including internally displaced persons (IDPs), returnees and refugees. This widespread and multilayered crisis impacts the social, political, and economic conditions of communities in Cameroon, Chad, the Niger and Nigeria. As humanitarian and development needs continue to grow and evolve, there is evidence of displaced persons returning to their areas of origin or habitual residence (DTM Nigeria, Cameroon and Chad). In 2018, IOM started documenting a growing number of displaced persons returning to their communities of origin in the Lake Chad Basin. In 2020, according to IOM, the Office of the United Nations High Commissioner for Refugees, and national government data, more than 1.7 million returnees (former IDPs) resided in Nigeria, Chad, Cameroon and the Niger. Approximately 34 per cent of the total displaced population identified by DTM in the Lake Chad Basin at the end of 2019 have returned to their communities of origin.

DTM's initial venture into stability and fragility measurements in Nigeria can be broken down into the five stages of the data science life cycle identified in DSEG's Framework, as displayed in Figure 2. IOM launched the Stability Index in Nigeria, Chad and Cameroon in 2019,[5] which measured levels of stability in areas that had been affected by conflict and displacement, to understand what makes returns more durable. The Stability Index measured returnees' perceptions of relative levels of stability and analysed which factors were more impactful on populations' decisions to remain in place or to move. To construct the index, a variety of metrics of stability were compared against key informants' perceptions of overall stability in a given area to determine the weight of each metric. The goal of the index was to conceptualize new and innovative ways to use data to better understand what stabilization means to returnees, and what factors influenced/supported returnees' decisions and made returns more sustainable. The tool was used to inform transition and recovery programming of IOM, partners, and governments to strengthen stability in regions affected by conflict, increased returns, or displacement.

[5] See: IOM, 2020.

Figure 3. Results from the Nigeria and Lake Chad Stability Index

The Stability Index measured returnees' perceptions of relative levels of stability, and analysed which factors were more impactful on populations' decisions to remain in place or to move.



Note: This map is for illustration purposes only. The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by the International Organization for Migration.

Stage 0: Fundamentals for Nigeria's Stability Index

Projection methods seek to reduce the number of dimensions in the data while also preserving the most important structure or relationships between the variables observed in the data. The first stage, *fundamentals*, refers more broadly to the internal ethical frameworks and guidelines held by the organization and the members of that organization undertaking the analytical effort. These systems of ethics include humanitarian principles, human rights, artificial intelligence ethics and humanitarian innovation.

Stage 1: Problem recognition and searching for solutions for Nigeria

The next stage is *problem solution and exploration*, in which a problem is identified, and project managers and data scientists clearly justify why data science is the most appropriate solution. In the case of Nigeria, the growing problem presented returning IDPs to their places of habitual residence, and an inability to understand or measure returnees' perceptions of the conditions of the area upon return. The "ideal" solution would help DTM identify returnees' perceptions of the stability of the locations to which they returned, and if returnees planned to stay or faced a secondary displacement. The search for solutions included a thorough understanding of the existing literature on the subject, as well as a robust exploration of the advantage of machine learning, or other advanced data science, over traditional solutions. For the Nigeria Stability Index, it was clear that using a machine-learning model was necessary to reduce noise in the data while also preserving the most important structure or relationships between the variables observed in the data. While considering all options available to understand this information gap, it was essential to keep in mind two core principles of ethical data science practices.

All humanitarian solutions should be needs-based, not technology-based. This means that the inclination to use data science should be driven by the needs of affected populations and the demands of a problem, not the reverse. Second, solutions should be driven by need and not by the allure of using high-tech solutions to attract donors for greater visibility or funding. Often the principle of Occam's razor applies here: the best solution is often the simplest one, not the most complex or highest tech.

All humanitarian solutions should be needs based, not technology based. This means that the inclination to use data science should be driven by the needs of affected populations and the demands of a problem, not the reverse.

For the Stability Index in Nigeria, the application of advanced data science methods was necessary in conjunction with the need to condense a large amount of information about a certain location into a digestible and interpretable value. This could then be used to give a reading of the situation on the ground to better inform on-the-ground planning and operations related to transitions and recovery. To support this type of social indexing for the Nigeria Stability Index, a complex analysis was vital. Given these considerations, the project proceeded to the next phase of the Ethical Framework.

Stage 2: Data collection in Nigeria

The second stage of the data science journey involves *data collection, data processing and protection*. For Nigeria, well-designed data collection was critical from an ethical standpoint. Given the complexity involved in capturing perceptions, phrasing questions and translation may influence comprehension. In this vein, perception surveys can unintentionally mask people's sentiments about certain topics as a result of how questions are phrased, or which translation techniques are used.

It was crucial for this project that distinct types of bias and their impact on the analysis were understood and clearly presented for transparency. One of the biggest areas of concern was the possibility of "response and representation bias", wherein data collected from human responses may not accurately represent the population. In this case, data collected by smartphone responses will capture only information from smartphone owners which (in humanitarian contexts) may exclude groups such as women, the elderly and lower-income people who may or may not have access to smartphones depending upon the context (IOM and SPIA, 2021). Any analysis based on this data will fundamentally misunderstand the affected population's characteristics that the analysis was aiming to measure. In Nigeria, as many female enumerators as possible were used to attempt to mitigate biases arising from this type of issue. However, the core concern was that key informants who responded to the survey about perceptions of stability were also a selected group, not necessarily representing the views of the general population. Taking this into account was necessary as the analysis progressed, especially since it impacted what types of indicators and measurements might further perpetuate this bias.

One of the core lessons learned from this exercise was the importance of language bias in affecting results. Since the translation of the survey into local languages was done ad hoc by enumerators, phrasing and questioning were inconsistent. The second round employed standardized translations to mitigate this, which ensured a more consistent survey method even when language barriers were encountered.

Stage 3: Data to processing – Applying algorithms to problems on the ground

Once the data-collection methods are understood and finalized, the project proceeds to the next stage – *data check* and *algorithm check*. This stage is critical in building on the foundation of the data collection and processing stage, but more notably, it has the potential to either conceal or reveal biases present from the beginning of the project.

Many organizations and industries have fluctuating definitions of algorithms, machine learning and even advanced data science, so it is useful to take a step back to define terms clearly and precisely. For the DSEG Ethical Framework and in this chapter, an algorithm can be defined as *a step-by-step, often automated mathematical and mechanical procedure for computing some mathematical function and/or solving a specific problem*. "The word 'algorithm' represents an abstract method for computing a function, while a program can be understood as an embodiment of a computational method in some programming language" (Dodgson et al., 2020).

Algorithms' automation power can be useful but can also alienate and obscure human input from processes that affect people. The use or overuse of algorithms can thus pose risks to populations affected by algorithm processes, as human input to such processes is often an essential element of protection or restitution for affected groups. Algorithms can deepen existing inequalities between people or groups and exacerbate the disenfranchisement of specific vulnerable demographics. More so than other types of data analysis, algorithms have the potential to create harmful feedback loops that can become tautological in nature and go unchecked due to the very nature of an algorithm's automation, in addition to raising concerns about the transparency of the model. In the humanitarian context, algorithms should be deployed only when they constitute a humanitarian information activity.[6] In other words, algorithms should be employed only in needs-based humanitarian situations that directly benefit the population in question. For the Nigeria Stability Index, it was critical to unequivocally link the perception-based survey results to clear programmatic initiatives and goals. This should be an essential component, part of meeting an identified community need, and should be deployed independently of State, private and external actor interests. Additionally, the use of algorithms should be neutral and impartial in accordance with humanitarian principles.

In the Stability Index in Nigeria, the decision-making process began with a categorization of ethical issues and concern. The first category was the technical rigour and properties of the algorithm and inputs. Considerations in this category included how to select the model based on existing statistical and empirical precedent for the method, selecting methods for algorithm validations, and being able to clearly identify the potential bias and limitations of the algorithm. This stage asks analysts to consider a few questions related to context and control of the inputs and outputs. In environments of partnerships, government relationships and donor relations, who controls what part of the algorithm and data life cycle? This question was less relevant to the Stability Index in Nigeria, as it was an internal DTM exercise, but it important to consider as the index expands to include other stakeholders.

Other important questions for consideration at this stage include: *Who has access to the data input and the output of the algorithm?* This considers the "leave-behinds" of the data and algorithm in the affected community. In addition, that the results of the algorithm or the code for the algorithm could potentially cause enhanced intelligence capacity for governments or other outside actors is important. In the context of Nigeria and the greater Lake Chad Basin, careful consideration was given to how the model was adapted from another context, and if the model would remain applicable outside of Nigeria.

Finally, it is also essential to conduct several rigorous reviews of the algorithm before deploying it. These reviews are both statistical and conceptual. For example, at this juncture, it is critical to identify key stakeholders affected by the algorithm and how potential algorithm failures might impact the most affected groups. This stage should also include an analysis of what it means to have false positives or false negatives in the model. In the case of Nigeria, it meant considering what the consequences might be if the team identified a location as "stable" when it really was not, or conversely, if a stable location was identified as "unstable". It is not difficult to imagine the consequence of neglecting programmes in an unstable area because the index misidentified it. Understanding the magnitude of these potential issues and how they interact is critical before undertaking any type of analysis.

[6] According to the *Signal Code*, humanitarian information activities (HIAs) are "[a]ctivities and programs which may include the collection, storage, processing, analysis, further use, transmission, and public release of data and other forms of information by humanitarian actors and/or affected communities. HIAs also include the establishment and development of communications capacity and infrastructure by responders and/or populations. These activities occur as part of humanitarian action throughout the response cycle and include, but are not limited to, improving situational awareness; disaster preparedness and mitigation; intervention design and evaluation; connecting populations to response activities and to each other; and supporting ongoing operations, including the delivery of assistance." (Greenwood et al., 2021)

Stage 4: Reliance on outputs – Communication, transparency and accountability in Nigeria

The final essential category of questions for consideration emerges at Stage 4 and can be thematically grouped around how to communicate the inputs, methods and results of the model to outside audiences. This is referred to as "communication, transparency and accountability" and includes the following questions:

(a) Is the algorithm actionable and relevant to the community/beneficiaries it serves?
(b) Is there clear communication of what was included and excluded from the model?
(c) How easy would it be for another data scientist to replicate the model?
(d) Is the code coherent and sharable? (Even if it is proprietary, internal sharing and peer review are essential.)
(e) Can affected populations meaningfully appeal decisions made about them?

In the case of Nigeria, this took the form of ground-truthing, and taking time after the analysis to check with relevant actors in each location and ensure that the results from the model reflected conditions and perceptions on the ground. This also meant taking time to build local capacities to replicate the model without the explicit presence of the main data scientist.

It is also useful here to examine the critical issue of "legibility" in the context of construction and implementation of any algorithm or model. Legibility refers to the understandability of the algorithm, or the ability of audiences with limited technical capacity to understand its mechanisms and outcomes and how it directly translates into humanitarian operations. Many ethical issues arise from a lack of legibility of what leads to the outputs of the algorithm. This should always be considered when weighing technical rigour and functional form against accessibility and legibility. Ultimately, algorithms and other machine learning-based models have the potential to help humanitarians solve problems more accurately and efficiently. They can be useful if and only if proper consideration is given to ethical implementation and harm reduction throughout the entire life cycle.

Conclusion

The many applications of data science projects to support humanitarian projects are complex and multifaceted, and would require careful thought regardless of ethical considerations. To be able to innovate responsibly, developers and practitioners must fully understand the problem at hand, prior to any model development. Without this, they run the risk of unnecessarily exposing displaced and mobile populations to additional risks and harms. Following a thorough process, such as the DSEG Ethical Framework, works to better position project teams to consider key questions, and understand key identified essential steps and necessary risks for any humanitarian data science project. While such a framework would never eliminate all risks, a thorough risk-awareness process allows teams to critically assess the project at the various stages of its life cycle, given confidence at some points, and highlight areas in need of improvement. The goal is to support teams to understand the risk and thus enable responsible innovation.

KEY SOURCES*

Dodgson, K., P. Hirani, R. Trigwell and G. Bueermann

2020 *A Framework for the Ethical Use of Advanced Data Science Methods in the Humanitarian Sector*. Data Science and Ethics Group. Available at www.migrationdataportal.org/resource/framework-ethical-use-advanced-data-science-methods-humanitarian-sector.

Greenwood, F., C. Howarth, D. Escudero Poole, N.A. Raymond and D.P. Scarnecchia

2021 *The Signal Code: A Human Rights Approach to Information During Crisis*. Harvard Humanitarian Initiative. Available at https://hhi.harvard.edu/publications/signal-code-human-rights-approach-information-during-crisis.

International Organization for Migration (IOM)

2020 *Stability Index – DTM Report: Measuring Perceptions of Stability in Nigeria and Cameroon*. Dakar. Available at https://displacement.iom.int/reports/stability-index-%E2%80%94-cameroon-and-nigeria-september-2019.

IOM and Princeton School of Public and International Affairs (SPIA)

2021 *Assessing the Use of Call Detail Records (CDR) for Monitoring Mobility and Displacement*. Available at www.migrationdataportal.org/resource/assessing-use-call-detail-records-cdr-monitoring-mobility-and-displacement.

McDonald, S.M.

2016 *Ebola: A Big Data Disaster – Privacy, Property, and the Law of Disaster Experimentation*. The Centre for Internet and Society, CIS Papers 2016.01. Available at http://cis-india.org/papers/ebola-a-big-data-disaster.

Raymond, N. and Z. Al-Achkar

2016 Building data responsibility into humanitarian action. United Nations Office for the Coordination of Humanitarian Affairs (OCHA) Policy and Studies Series, 018. Available at https://thegovlab.org/static/files/publications/TB18_Data%20Responsibility_Online.pdf.

** All hyperlinks were working at the time of writing this report.*

REGULATORY FRAMEWORKS FOR CROSS-SECTORAL DATA PARTNERSHIPS

UNDERSTANDING POPULATION MOVEMENT UNDER UNCERTAINTY IN HUMANITARIAN AND PUBLIC HEALTH EMERGENCIES

Miguel Luengo-Oroz,[1] *Katherine Hoffmann Pham*[2] *and Rebeca Moreno Jiménez*[3]

Introduction

Humanitarian agencies must be prepared to mobilize quickly in response to complex emergencies, and their effectiveness depends on their ability to identify, anticipate and prepare for future needs. To support preparedness planning, there has been growing interest in the use of novel predictive modelling tools and data sources to forecast displacement. For example, predictive models have been developed for specific displacement settings such as Somalia, Yemen, the Syrian Arab Republic, Mali, Burundi and the Central African Republic (Earney and Moreno Jimenez, 2019; Huynh and Basu, 2020; Suleimenova et al., 2017), as well as at a global scale (Kjaerum, 2020). More generally, approaches range from short-term early warning systems to midterm model- or survey-based forecasts and long-term foresight exercises (Sohst and Tjaden, 2020).

However, the COVID-19 pandemic introduced new challenges for prediction by creating large-scale, sudden and unexpected global changes in mobility. Lockdowns, border closures, the economic impacts of COVID-19, and individual behavioral changes have all influenced travel patterns. This chapter discusses the development of predictive modelling and situational awareness tools in the face of high levels of uncertainty.

Specifically, it describes a three-part decision support system designed to help the Office of the United Nations High Commissioner for Refugees (UNHCR) estimate, model and anticipate the movements of displaced people from the Bolivarian Republic of Venezuela

[1] Miguel Luengo-Oroz is Senior Advisor on Frontier Technologies at the United Nations Global Pulse, at the Executive Office of the United Nations Secretary-General. He is an expert in data science for humanitarian response and global health. He is a professor at the Universidad Politécnica de Madrid (UPM). He holds a PhD and MScEng from UPM and an MSc from the Ecole des Hautes Études en Sciences Sociales in Paris.

[2] Katherine Hoffmann Pham is an Artificial Intelligence Researcher at the United Nations Global Pulse. Previously Katherine worked at the Innovation for Poverty Action for almost four years. She holds a master's in International Policy Studies from Stanford University and is a PhD candidate at New York University.

[3] Rebeca Moreno Jiménez is Innovation Officer and Data Scientist of the UNHCR's Innovation Service. Rebeca holds a bachelor's degree in International Relations and an MPP from the Instituto Tecnológico y de Estudios Superiores de Monterrey, and an MPA focusing on the management of technology policy and humanitarian affairs from Columbia University.

to Brazil in the aftermath of COVID-19-related border closures. The initial goal[4] of the project was to support the contingency planning process by evaluating the likelihood of different displacement scenarios using innovative data sources and methods. The system consists of the following:

(a) **An interactive simulation tool** to model how people move through the different stages of border crossing under various assumptions;
(b) **Collection of real-time "nowcast" data** from diverse sources (e.g. Google, the Armed Conflict Location and Event Data Project (ACLED), radio) which might identify large numbers of people travelling to (or planning to travel to) the border;
(c) **Mathematical models** to estimate future arrival levels according to different economic and social indicators.

Such evidence-based decision support tools help UNHCR to advocate the provision of necessary resources such as shelter, medical services, COVID-19 isolation areas, vaccine delivery, and access to documentation for undocumented migrants in order to connect them to aid and public services.

As noted above, the core modelling challenge in this setting is that there are large periods of unusual arrivals data as a result of the closure of the Brazil–Bolivarian Republic of Venezuela border in one or both directions due to the pandemic. Since the situation is constantly changing, and many Venezuelans face mobility restrictions, current patterns are unlikely to be representative of future mobility, and there is no clear expectation about how many people will cross when given the opportunity to do so. Therefore, traditional predictive models are likely to be inaccurate. The system's three-part approach reflects an attempt to address this uncertainty by adopting multiple strategies to anticipate future movements.

In the remainder of this chapter, we provide context on the problem statement and describe the three components of the predictive modelling and situational awareness system. While the system has been developed for a specific border-crossing setting, its components and the overall approach could easily be adapted to other displacement settings. This is an area of ongoing work. In order to guide practitioners interested in developing similar systems, this chapter concludes with a discussion of challenges and lessons learned regarding partnerships, data, risks and harms, and model validation.

Context

The Bolivarian Republic of Venezuela is currently experiencing one of the largest contemporary displacement crises: an estimated 5.4 million Venezuelans live abroad (UNHCR, n.d.), and Venezuelans represent almost 20 per cent of asylum seekers globally (UNHCR, 2021). While Colombia is the primary destination of displaced Venezuelans (R4V, 2021; Otis, 2021; Frydenlund et al., 2021), as of 2019 over 220,000 Venezuelans resided in Brazil (UNHCR, 2019), and Venezuelans have historically moved regularly between these countries for protection, economic or family reasons. However, on 18 March 2020, the land border between the Bolivarian Republic of Venezuela and Brazil was closed in response to the COVID-19 pandemic.

[4] As the border has partially reopened, the situation has moved beyond the scenarios envisioned in the initial contingency plan. Consequently, each component of the system is undergoing continual updates to reflect new data, new realities on the ground, and new modelling needs.

In anticipation of a possible border reopening in 2021, the operational team at the border (UNHCR Brazil, Boa Vista sub-office) requested support from remote data science teams (United Nations Global Pulse, UNHCR Global Data Service and UNHCR Innovation Service) to estimate the number of displaced persons expected to enter Brazil from the Bolivarian Republic of Venezuela, both during and after the official border closure. This information was needed as an input to the contingency planning process, to determine whether existing shelter capacity should be expanded in anticipation of these new movements.

A fundamental challenge in the contingency planning process is the uncertainty around anticipated arrivals. The number of people who will cross the border upon reopening is expected to be a function of two key factors. The first is the baseline number of aspiring migrants and refugees who would have crossed the border with or without the pandemic but were prevented by the closure. For example, these may include people who wish to move for economic reasons, or who have plans to join their family in Brazil. The second is the impact of COVID-19, which may have increased or decreased intentions to cross the border. For example, COVID-19 may increase border crossings by driving Venezuelans to seek access to medical supplies, facilities and care in Brazil. On the other hand, the economic impacts of COVID-19 may decrease the appeal of crossing by reducing the number of job opportunities for Venezuelans living abroad.

Tools and models for operational support

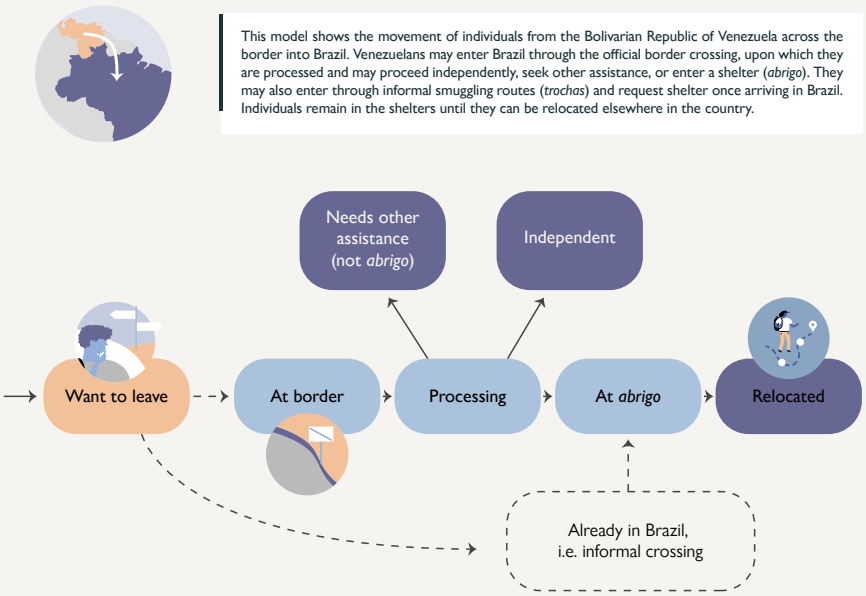
The decision support system consists of three primary tools to inform the contingency planning process. The first is an interactive simulation tool for arrivals. This tool does not require precise forecasts of the number of displaced persons, but rather it allows decision makers to experiment with assumptions about the rate of new arrivals and shelter capacity. The second is a "nowcast" data set of political, economic, mobility, health and other indicators to capture factors related to border-crossing intentions. It integrates official and unofficial data sets from a wide variety of sources, many of which are also available for other displacement settings. The third is a set of mathematical models for predicting arrivals based on past border crossings. While these models produce displacement estimates, these estimates are subject to uncertainty in the input data and in the model construction and formulation. Therefore, all three components of this system are meant to act as complements to each other, with the mathematical models generating forecasts of potential outcomes, the simulation tool encouraging decision makers to be prepared for different possible scenarios, and the nowcast data sets helping decision makers to monitor emergent trends.

Simulation tool for border crossing

The first component of the system is an interactive simulation tool for arrivals.[5] The conceptual map of the border-crossing process, which was developed through consultations with the Roraima team and is implemented by the tool, is shown in Figure 1. A typical journey includes five key steps: wanting to leave, travelling to the border, formal processing at the border crossing, moving into an *abrigo* (shelter) and relocating into the interior of Brazil. The model works by performing a simulation that moves individuals from one step to the next at different rates until they are relocated or otherwise exit the model.

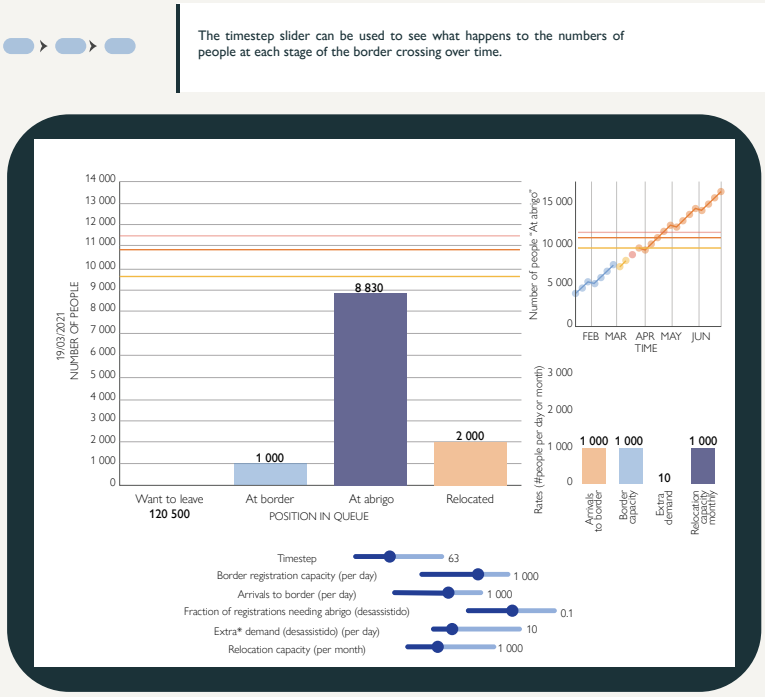
[5] Available at https://brazil-venezuela-flows.unglobalpulse.net/.

Figure 1. Conceptual model of the border-crossing process



Operational teams can use sliders to adjust assumptions about arrivals and registration rates, the need for shelter among new arrivals, and the rate at which displaced people can be relocated from shelters into the interior of Brazil. Based on these assumptions, the tool projects the number of people in shelters, and notes the dates on which different capacity limits are reached (triggering an operational response). The time step slider can be used to see what happens to the numbers of people at each stage of the border crossing over time. This tool makes it possible to conduct a sensitivity analysis to explore how quickly different contingency planning phases are triggered, based on varying assumptions about arrivals and capacity.

Figure 2. A snapshot of the interactive simulation tool for arrivals



Real-time "nowcast" data

The second component of the system is a set of data on displacement proxies, drivers, and correlates from a variety of quantitative and qualitative sources. The focus is primarily on data from the Bolivarian Republic of Venezuela, since volatility amid the national economic and political situation is a key driver of displacement from the country (van Praag, 2019), and the population of interest consists of people currently within the Bolivarian Republic of Venezuela who intend to cross the border. However, many of the sources described below are cross-national data sets that can be collected for multiple countries, and they can be included into such models to help understand "pull" and "push" factors. In particular, information has been collected from the following:

(a) **Operational data sets.** These were provided by UNHCR and include registrations of new arrivals at the border between the Bolivarian Republic of Venezuela and Brazil, historical data on official daily crossings going back to 2018, data on current and potential shelter capacity, and data from surveys assessing the needs of displaced Venezuelans.

(b) **Data on conflicts and protests.** Data on unrest are collected from the Armed Conflict Location and Event Data set,[6] which contains records of violent and nonviolent incidents around the world (Raleigh et al., 2010). In addition to gathering aggregate data on fatality and incident counts, incident descriptions are parsed in order to categorize events as related to fuel, food security, the economy, wages or working conditions, health or medicine, and utilities (i.e. electricity and water).

(c) **Internet search data.** Google Trends[7] is used to capture trends in the volume of Google searches originating from the Bolivarian Republic of Venezuela. Trends are created for topics such as "Brazil–Venezuela border", cities/municipalities such as "Pacaraima" and specific keywords such as "trochas" (which refer to pathways for crossing the border irregularly).

(d) **Social media data.** An out-of-the-box social listening tool is used to track trends in the mentions of different keywords (e.g. "Frontera de Brasil", "trocha/trochero") over time. Furthermore, a query taxonomy has been developed to identify social media conversations related to border crossings, which are then analysed to find emerging stories and topics of interest.

(e) **Mobility data.** Google's COVID-19 Community Mobility Reports[8] are used to track the aggregate volume of visits to different venue categories (i.e. workplaces, residences and transit stations) over time, capturing internal restrictions on movement in the Bolivarian Republic of Venezuela. Furthermore, Facebook's Marketing API[9] is used to obtain estimates of the number of daily and monthly active Facebook users in towns along the border-crossing route, according to whether that location is the user's home location, one of their recent locations, or a location they are visiting while travelling.[10] On the Brazilian side of the border, it is also possible to obtain estimates of the number of people who formerly lived in the Bolivarian Republic of Venezuela (and vice versa). By querying the API repeatedly over time, it is possible to collect time-series data on how these populations vary and to monitor changes in trends.

[6] Available at http://acleddata.com.

[7] Available at https://trends.google.com/.

[8] Available at www.google.com/covid19/mobility/.

[9] Available at https://developers.facebook.com/docs/marketing-apis/.

[10] "Home" locations refer to users' self-declared home cities, which Facebook attempts to confirm by checking each user's IP address and the location of their friends. Facebook defines "travelled in" locations as recent locations which are over 100 miles from the user's "home" location. Full details can be found in the Facebook Marketing API's documentation, available at https://developers.facebook.com/docs/marketing-api/audiences/reference/basic-targeting#location.

(f) **Other data sets.** Situation-specific data have also been gathered from a variety of other sources, including data on COVID-19 cases and fatality rates,[11] COVID-19 self-reported symptoms,[12] oil prices,[13] consumer price indices,[14] exchange rates,[15] radio transcripts and satellite data.

The compiled and aggregated data sets are used for two primary purposes. On the one hand, the raw data can be presented directly to the team in Roraima to provide real-time insights into displacement drivers and trends, and information on the demographics of informal crossings which may not be recorded by official statistics. For this purpose, data on social media conversations and Facebook audience estimates were particularly useful, as they could be compared with observations and experiences on the ground. On the other hand, the raw data can be used as an input into predictive models, in order to produce forecasts and sensitivity analyses. For this purpose, Google Trends and the Google Community Mobility Reports[16] provided several useful predictive features.

Mathematical models for predicting arrivals

As noted above, while the arrivals simulation tool is designed to explore shelter capacity and policy responses as a function of different assumptions about arrival rates, it is difficult to anticipate what the rate of arrivals will be when the border does reopen. Therefore, the third component of the system includes experimental time-series forecasting strategies to predict arrivals using different characteristics of the political, economic and COVID-19 situation in the Bolivarian Republic of Venezuela. These forecasting models assume that arrivals are not random, but rather they are driven by factors such as political unrest and economic well-being, which can be captured by quantifiable data.

The modelling approach consists of three main steps. The first is the creation of a standardized data set with indicators that are expected to be relevant to border-crossing demand. These include variables from the data sets described above, as well as manually coded variables to capture the schedules of buses which carry people to the border, school holidays and key historical events.

In the second step, a number of different time-series forecasting models are fitted to predict arrivals. These include gravity models, long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), and common machine-learning models like lasso (Tibshirani, 1996) and ridge (Hoerl and Kennard, 1970) regressions, random forests (Breiman, 2001), AdaBoost (Freund and Schapire, 1997), and XGBoost (Chen and Guestrin, 2016).[17] In general, models were set up to predict

[11] Available at https://ourworldindata.org/coronavirus.

[12] Available at https://gisumd.github.io/COVID-19-API-Documentation/.

[13] Available at www.eia.gov/dnav/pet/PET_PRI_SPT_S1_D.htm.

[14] Available at www.fao.org/faostat/en/#data/CP/metadata.

[15] Available at www.investing.com/currencies/usd-vef.

[16] Note that Google Community Mobility Reports have only been made available since the start of the pandemic, so they have been used only for models trained on informal crossings during the pandemic.

[17] Gravity models focus on predicting pairwise flows (e.g. from the Bolivarian Republic of Venezuela to Brazil) using a linear combination of features of the origin and/or destination, features of the origin–destination pair, and other factors (such as time) (Ramos, 2016). LSTMs are a type of neural network that are specifically designed for sequence modelling, and which can "carry forward" information from earlier points in time (i.e. they have a "memory") (Nicholson, 2020). The standard machine-learning models listed above vary in their approaches. Lasso and ridge regressions predict arrivals in Brazil as a linear combination of the input features (e.g. Google searches for the topic "Brazil–Venezuela border") with the addition of different penalty functions to aid in focusing the model on the most important features. Random forests, AdaBoost and XGBoost are methods that rely on decision trees, which divide up the input data set into similar subsets based on its features and then make a prediction for each subset. A helpful introduction to many standard machine-learning algorithms can be found in the work of James et al. (2013).

daily[18] arrivals 30 days in advance. The models were initially trained and validated on data prior to the border closing, in order to produce forecasts for the period of the border closure. More recently, models have been trained using data on requests for shelter registered by UNHCR during the pandemic (which generally result from informal crossings), in order to forecast arrivals prior to the full reopening of the border.

The third step involves estimating uncertainty around the model predictions and comparing the results of different models. The system focuses on model uncertainty, which can result from sensitivity to the exact data points used to train a machine-learning algorithm.[19,20] By comparing the results from different modelling approaches and their associated prediction intervals, it is possible to evaluate the likelihood of different arrival scenarios and observe which scenarios are consistent with model predictions.

Discussion and lessons learned

This section discusses some of the highlights, risks and challenges of this work. In particular, it addresses the benefits of co-design involving operational and data science teams; the challenges of gathering and integrating heterogeneous data sets; the potential risks and harms of forecasting mobility; and the difficulty of validating predictive models in unprecedented situations with high uncertainty, such as the COVID-19 pandemic.

Partnerships

A key highlight of this project has been the collaboration between stakeholders on the ground in Roraima and the remote data science teams. Given the political and operational complexities of the context, iterative co-design has been critical to the project. As the situation on the ground evolves, the project deliverables and objectives have been continuously adapted; for example, in early 2021 the project shifted to focus on better understanding *informal* crossings, since there was a surge in demand for shelter even as the border remained formally closed until June 2021.

One of the unique aspects of the data-collection effort was the speed with which diverse sets of data were aggregated and analysed: within one month of the initial request from UNHCR Brazil's Boa Vista sub-office, it was possible to produce data-driven reports analysing the situation in the Bolivarian Republic of Venezuela. Rapid mobilization was enabled by two key factors. First, the collaboration began with a project-scoping exercise, which aimed to define the modelling problem and conduct a census of available data sources. This allowed the team to quickly identify information gaps and needs. Second, the data-collection effort benefited from United Nations Global Pulse's and UNHCR's previous experience working with many of the target data sources. Because data-collection pipelines can be time-consuming to set up, it is worthwhile to invest in generalizable processes for accessing

[18] With the exception of the gravity model, which was trained on monthly data.

[19] Model uncertainty can also result from random choices made by the model algorithm, known as stochasticity (for example, some models sample a random subset of features or data points at different stages of the training process), but the method currently used to compute prediction intervals is designed to address the former source of model uncertainty. Furthermore, we note that there may be uncertainty in the training data itself, due to variations in data quality and accuracy.

[20] A bootstrapping approach (Nielsen, 2020) is adopted to quantify uncertainty for the standard machine-learning models listed above. For a fixed number of iterations, a subset of the training data is randomly selected, and a model is fit to this data. A prediction is then produced on the remaining observations which were not used to fit the model. As a result, numerous models are produced which have been fitted on different samples from the training data set, each with slightly different predictions. This allows for the calculation of prediction intervals for the forecasts produced by each type of model. Prediction intervals are calculated at the 95 per cent confidence level.

and analysing cross-national data sets, which make it possible to quickly adapt existing code and pipelines to new challenges.

Data

The different data sources discussed in Section 3.2 ("Real-time 'nowcast' data") come with distinct advantages and challenges, which are often associated with the format in which the data set is provided. Some considerations encountered with common data access formats include the following:

(g) **Application programming interfaces (APIs).** APIs are standardized interfaces created by data providers (e.g. companies such as Google) in order to allow others to retrieve data of interest; while they are typically easy to access in real time, they often require programming expertise to use, and may impose rate limits on the amount of data requested.

(h) **Off-the-shelf tools.** Off-the-shelf tools (e.g. Talkwalker,[21] Dataminr,[22] NetBase Quid[23] and Brandwatch[24]) are particularly popular for social media and media analysis, and act as intermediaries (data brokers) between data providers and end users; they are generally intuitive to use, but may require a partnership or license agreement to access and may act as "black box" tools which provide little insight into how analysis is conducted.

(i) **Custom pipelines.** For less conventional sources of data such as radio, it may be necessary to build a custom pipeline to gather and process data; while this process can be time- and labour-intensive, it offers the most control over data collection and analysis.

In addition to these format-specific challenges, another challenge involves variations in time frequency, geographic precision, and update speed across data sets. For example, it was often necessary to aggregate daily data or resample monthly data in order to create indicators that could be compared on a weekly basis. The usefulness of data for nowcasting and forecasting purposes also depends on whether the data sets are updated regularly, or if they are available only with a time lag.

A further concern is uncertainty in the data sources themselves. For example, data on official border crossings do not include those who cross informally through *trochas*, making it difficult to know the true population that is potentially in need of shelter. Data on COVID-19 case counts and fatalities may be affected by the Venezuelan Government's interference to prevent accurate reporting of the statistics (Cincurova, 2020). Aggregation and resampling, as described above, may introduce additional variation which can contribute to increasing the uncertainty of models fitted on this data.

A final challenge is the "black box" nature of the data offered by large online platforms. For example, while Facebook audience estimates can be used to study changes in migrant *stocks* over time, it is not possible to track *flows* from one city to another using the Marketing API. Furthermore, it is not always clear how concepts like "formerly lived in Venezuela" are encoded by the platform, and the platform's definition of terms such as "active users" (or the algorithms that estimate their numbers) may change over time. While these data sets may still be useful for prediction, the data should be interpreted with caution and an awareness of their limitations.

[21] Available at www.talkwalker.com/.

[22] Available at www.dataminr.com/.

[23] Available at https://netbasequid.com/.

[24] Available at www.brandwatch.com/.

Risks and harms

As with all predictive analytics projects, it is important to understand how predictions will be used and to consider potential risks and harms. In this case, predicting arrivals is clearly aligned with UNHCR's mandate to provide services to persons of concern, and having such predictions can help UNHCR advocate on behalf of these populations. Nevertheless, some findings of the data analysis might be sensitive and have potential for harm by malicious actors, and so should best be kept internal to the operation.

In addition to these broader concerns, there are risks and harms associated with each of the individual data sets used. For example, social media data overrepresent those with access to the Internet, as well as users whose voices are amplified by platform algorithms. Radio data can raise privacy concerns, because while the data are technically produced for public consumption, participants on radio shows may reveal personal information about themselves without realizing that their voices and stories are being captured and studied by analysts thousands of miles away. Although data sets may be accessible through public websites, pre-existing access agreements, or procurement, it is important to follow required due diligence processes and assess the potential risks and harms of individual data sets and their combinations. Even internally collected data carry risks because they may contain sensitive information on populations in vulnerable situations that could be targeted by governments or adversarial actors. It is important to undertake mitigation actions to reduce all of the risks previously identified. For example, any personally identifiable information that is not necessary for analysis should be removed from data sets even prior to sharing *within* an organization.

Model validation

As noted throughout this contribution, the Bolivarian Republic of Venezuela–Brazil border reopening represents a particularly challenging setting for forecasting because of the lack of representative, recent arrivals data due to the border closure and high uncertainty from multiple sources. One limitation of the initial mathematical forecasting approach was that it did not explicitly incorporate adjustments for the COVID-19 pandemic, because of the large uncertainty around how COVID-19 affects intentions to cross the border. Rather, it predicted arrivals given economic and social conditions, using models fitted on data from before the border closure. The effects of the pandemic may have entered the models indirectly when predicting arrivals statistics, but only to the extent that they influence input variables such as consumer price indices or the number of protest events in the Bolivarian Republic of Venezuela.

Given predictions based on pre-COVID-19 trends, it is possible to scale these predictions using assumptions about how COVID-19 might have changed migration intentions, or refit models using more recent data. One challenge is that these assumptions cannot be validated until the border itself fully reopens. For this reason, we feel that the arrivals simulation tool is particularly important, because it encourages experimentation with different assumptions and avoids framing arrival forecasts in a deterministic way.

Conclusion

This chapter highlights some of the opportunities and challenges in modelling population movement during the COVID-19 pandemic – and in uncertain situations more broadly. The pandemic has created new policy needs that require effective operational responses. At the same time, the pandemic has also generated a great deal of unprecedented behavior, which makes forecasting difficult.

One innovative aspect of this project is that predictive models are complemented by a simulation tool for arrivals and a data set of "nowcast" indicators. This three-part approach was developed based on an understanding that traditional forecasting models may be inaccurate at predicting future border crossings. The simulation tool encourages decision makers to experiment with assumptions and consider different possible arrival rates and outcomes, rather than relying on a fixed set of predictions, and leverages information about possible futures generated by the prediction models. At the same time, rapid data collection from a variety of sources facilitates an understanding of the situation on the ground in an evolving crisis. While mathematical models are used to make predictions, these models are also accompanied by prediction intervals to highlight ongoing uncertainty to decision makers.

A key lesson learned from this case study is that field and data science teams should be open to a wide range of data sources and methodologies, adopting a diverse approach which can avoid some of the shortcomings that arise from considering only a narrow set of data sources or a single modelling approach. Aware that individual data sources may have biases or limitations – for example, the number of protests during the pandemic may be artificially reduced due to lockdowns, and the number of Google searches for information about the border may spike because of interest in crossing *or* because a closure is announced – this project sought to gather as many different sources of data as possible and to analyse them from different perspectives. The approaches described here represent work in progress, and project goals and deliverables are continually modified as the situation on the ground evolves. Such flexible multi-method approaches are critical to providing the agility needed to respond to this pandemic and future humanitarian emergencies.

REFERENCES*

Breiman, L.
2001 Random forests. *Machine Learning*, 45(1):5–32.

Chen, T. and C. Guestrin
2016 XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–94.

Cincurova, S.
2020 Known unknowns: The challenge of collecting COVID-19 data in Venezuela. *The New Humanitarian*, 4 June. Available at www.thenewhumanitarian.org/news-feature/2020/06/04/coronavirus-data-Venezuela.

Earney, C. and R. Moreno Jimenez
2019 Pioneering predictive analytics for decision-making in forced displacement contexts. In: *Guide to Mobile Data Analytics in Refugee Scenarios* (Salah, A.A., A. Pentland, B. Lepri and E. Letouzé, eds.). Springer Nature Switzerland AG, Cham, pp. 101–119.

Freund, Y. and R.E. Schapire
1997 A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Frydenlund, E., J.J. Padilla and K. Palacio
2021 Colombia gives nearly 1 million Venezuelan migrants legal status and right to work. *The Conversation*, 14 April. Available at http://theconversation.com/colombia-gives-nearly-1-million-venezuelan-migrants-legal-status-and-right-to-work-155448.

Hochreiter, S. and J. Schmidhuber
1997 Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hoerl, A.E. and R.W. Kennard
1970 Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huynh, B.Q. and S. Basu
2020 Forecasting internally displaced population migration patterns in Syria and Yemen. *Disaster Medicine and Public Health Preparedness*, 14(3):302–307.

** All hyperlinks were working at the time of writing this report.*

Inter-Agency Coordination Platform for Refugees and Migrants from Venezuela (R4V)

2021 R4V Latin America and the Caribbean: Venezuelan refugees and migrants in the region – September 2021. Available at www.r4v.info/en/document/r4v-latin-america-and-caribbean-venezuelan-refugees-and-migrants-region-september2021.

James, G., D. Witten, T. Hastie and R. Tibshirani

2013 *Introduction to Statistical Learning*. Springer. Available at www-bcf.usc.edu/~gareth/ISL/.

Kjaerum, A.

2020 Foresight: Using machine learning to forecast and understand forced displacement. *Migration Policy Practice*, X(4):26–30. Available at https://publications.iom.int/books/migration-policy-practice-vol-x-number-4-september-december-2020.

Nicholson, C.

2020 A beginner's guide to LSTMs and recurrent neural networks. Pathmind. Available at http://wiki.pathmind.com/lstm.

Nielsen, D.S.

2020 Bootstrapping prediction intervals. 1 March. Available at https://saattrupdan.github.io/2020-03-01-bootstrap-prediction/.

Office of the United Nations High Commissioner for Refugees (UNHCR)

2019 UNHCR welcomes Brazil's decision to recognize thousands of Venezuelans as refugees. 6 December. Available at www.unhcr.org/news/briefing/2019/12/5dea19f34/unhcr-welcomes-brazils-decision-recognize-thousands-venezuelans-refugees.html.

2021 Refugee Data Finder. Available at www.unhcr.org/refugee-statistics/.

n.d. Venezuela situation. Available at www.unhcr.org/venezuela-emergency.html.

Otis, J.

2021 "A huge opportunity": Venezuelan migrants welcome Colombia's new open-door policy. NPR, 26 February. Available at www.npr.org/2021/02/26/971776007/a-huge-opportunity-venezuelan-migrants-welcome-colombias-new-open-door-policy.

Raleigh, C., A. Linke, H. Hegre and J. Karlsen

2010 Introducing ACLED: An armed conflict location and event dataset – Special data feature. *Journal of Peace Research*, 47(5):651–660.

Ramos, R.

2016 Gravity models: A tool for migration analysis. IZA World of Labor. Available at https://doi.org/10.15185/izawol.239.

Sohst, R.R. and J. Tjaden

2020 Forecasting migration: A policy guide to common approaches and models. *Migration Policy Practice*, X(4):8–13. Available at https://publications.iom.int/books/migration-policy-practice-vol-x-number-4-september-december-2020.

Suleimenova, D., D. Bell and D. Groen

2017 A generalized simulation development approach for predicting refugee destinations. *Scientific Reports*, 7(1):13377.

Tibshirani, R.

1996 Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

van Praag, O.

2019 Understanding the Venezuelan refugee crisis [blog]. The Wilson Center, 13 September. Available at www.wilsoncenter.org/article/understanding-the-venezuelan-refugee-crisis.

15

MOBILE PHONES

TACKLING COVID-19

BUILDING DATA PARTNERSHIPS

DISPLAYING SOCIAL CONNECTIONS

MOBILITY EVIDENCE FROM MOBILE NETWORK OPERATORS DATA TO SUPPORT COVID-19 RESPONSE

Michele Vespe,[1] *Stefano Maria Iacus,*[2] *Umberto Minora,*[3] *Carlos Santamaria,*[4] *Francesco Sermi,*[5] *Spyridon Spyratos*[6] *and Dario Tarchi*[7]

Introduction

Mobility interventions have been adopted globally at the national and regional scale to reduce COVID-19 transmission. Stay-at-home recommendations, closure of public spaces or lockdown measures highlighted the need for data and solid evidence on the level of mobility, to quantify their effectiveness and better outline de-escalation and re-escalation strategies. Timely and granular mobility insights can be extracted from aggregate and anonymized mobile network operator (MNO) data. This chapter introduces a unique business-to-government (B2G) initiative between the European Commission and several MNOs in Europe to help fight the COVID-19 pandemic. Results that are comparable across countries include near real-time indicators of mobility, connectivity and highly interconnected regions such as mobility functional areas (MFAs). MFAs can provide data-driven input to tailored measures that would ensure a balance between epidemiological effects and socioeconomic impacts. In addition to scientific challenges associated with the processing of heterogeneous data from multiple private-sector entities, the initiative also highlighted aspects in the areas of security, privacy, fundamental rights and commercial confidentiality.

[1] Michele Vespe is Team Leader at the European Commission's Joint Research Centre (JRC). He leads the activities of researchers investigating societal consequences associated with the improved availability of digital trace data. This includes research in the fields of data governance and computational social science.

[2] Stefano Maria Iacus is an Officer at the JRC of the European Commission, formerly a Full Professor of Statistics at the University of Milan. He was an R-core member (1999–2017) and is an R Foundation member. His fields of expertise are migration, demography, computational statistics, simulation and inference for stochastic processes, causal inference, textual analysis and numerical finance.

[3] Umberto Minora works as a Researcher at the Knowledge Centre on Migration and Demography (KCMD) of the European Commission. He graduated with a PhD from the Department of Earth Sciences "Ardito Desio" of the University of Milan.

[4] Carlos Santamaria worked as Scientific Project Officer at the KCMD of the European Commission. He holds an MBA from the Universidad Politécnica de Madrid, Spain.

[5] Francesco Sermi is Scientific/Technical Project Officer at the KCMD of the European Commission. He holds a PhD in Informatics, Multimedia and Telecommunications Engineering from the University of Florence, Italy.

[6] Spyridon Spyratos is a Data Scientist at the European Union Intellectual Property Office. Prior to this position, he was a Researcher at the Knowledge Centre on Migration and Demography of the European Commission. Spyridon holds a PhD on big data, crowdsourcing and urban planning from the University of Thessaly, Greece.

[7] Dario Tarchi is Deputy Head of the Demography, Migration and Governance Unit at the European Commission's JRC. He is a physicist by training and has been involved, as a Project Leader, in the design, creation and operation of the KCMD since its establishment in 2016.

B2G data-sharing mechanisms are expected to significantly contribute to different stages of policy cycles and enhance situational awareness for emergency and policy responses in several fields. Despite clear limitations, digital trace data offer near real-time and granular signals over dimensions often not fully captured by official statistics, such as migration (Rango and Vespe, 2017 and 2021; Spyratos et al., 2018 and 2019). The potential impact of digital trace data is even more evident in domains where high resolution and timely information on mobility are of paramount importance, as during the first waves of the COVID-19 pandemic, when authorities and administrations had to rapidly react globally to limit the spread of outbreaks while reducing the pressure on national health systems. Spatial non-pharmaceutical interventions to contain mobility have been undertaken in different forms, from full lockdowns to more relaxed measures aiming at reducing contacts and therefore infections.

This chapter offers an overview of the main results of both extracting systematic evidence for policy and highlighting B2G data-sharing process features that may support similar future initiatives.

Key insights for governance models ensuring ethical and efficient data-sharing between partners across private and public sectors

Identifying and defining institutional requirements

Feeding epidemiological metapopulation models and understanding the impact of such measures on mobility, as well as the role of mobility during the first phases of the epidemic, were some of the uses identified in early April 2020, when the European Commission asked European MNOs to share aggregate and anonymized mobile positioning data. The terms of cooperation between MNOs and the European Commission are outlined by a Letter of Intent (GSMA, 2021). Insights on mobility patterns of population groups were meant to serve the following purposes in the fight against COVID-19:

(a) "understand the spatial dynamics of the epidemics using historical matrices of mobility national and international flows;
(b) quantify the impact of physical distancing measures (travel limitations, non-essential activities closures, total lock-down, etc.) on mobility ... ;
(c) feed [susceptible-infectious-recovered (SIR)] epidemiological models, contributing to the evaluation of the effects of physical distancing measures on the reduction of the rate of virus spread in terms of reproduction number (expected number of secondary cases generated by one case);
(d) feed models to estimate the economic costs of the different interventions, as well as the impact of control extended measures on intra-EU cross border flows [and traffic jams] due to the epidemic;
(e) [cover all member States in order to acquire insights relevant to COVID-19 from the entirety of the EU]."

Starting in April, data from European MNOs were gradually shared pro bono with the European Commission, covering 22 European Union member States plus Norway, on a daily basis, with an average latency of a few days, and in most cases covering historical data since February 2020. The data set enables comparative mobility analysis across countries. Different from openly available mobility

data derived, for instance, from online services[8] or mobile apps,[9] MNO data can be processed to provide insights on human mobility at a level of granularity (MNO data reach up to the municipal level while available app data are limited to the regional or provincial level), timeliness, frequency of update (some MNOs provide new updates several times a day), representativity (coverage of a large fraction of the population and almost all population groups) and transparency (see a more detailed methodological description below) that make this data set uniquely positioned to continuously improve the response to an evolving emergency.

Implementing safeguards to protect the fundamental rights of data subjects

The aspects of security and integrity of the data need to be immediately addressed by implementing end-to-end encryption protocols for data transmission, and by developing a secure dedicated platform to store and process the data, which are ultimately made accessible to a limited number of users. Although the data shared by MNOs (in the form of origin–destination matrices (ODMs)) contain only anonymized and aggregate data, a "reasonability test" was run upon the reception of preliminary data samples. This process served to verify that the specification in terms of origin–destination aggregate data was respected and to assess whether or not the risk of reidentification of the individuals was reasonably low. In addition, purpose limitation on the use of the data, access control and data retention horizon were some of the safeguards that helped address the legitimate business interests of operators and guarantee respect of fundamental rights. In addition, communication aspects had to be analysed in detail in order to appropriately ensure that the initiative involved anonymized and aggregate data only, and in order to avoid political backlash in such a sensitive domain. This required continuous consultation with all stakeholders involved in the initiative prior the publication of communication outlets or scientific results to the public.

Understanding the data and ensuring quality standards

Understanding the data in terms of what they capture, how they are collected, and their strengths and limitations is of utmost importance. In compliance with the Guidelines on the use of location data and contact tracing tools in the context of the COVID-19 outbreak by the European Data Protection Board (2020), MNO data shared in the framework of this initiative do not provide information about the behaviour of individuals. Nevertheless, the data can offer useful insights into the mobility patterns of various population groups. The data hosted and processed are in the form of ODMs (Mamei et al., 2019). Each *ODMi,j* cell of the matrix shows the overall number of *movements* (also referred to as *trips* or *visits*) that have been recorded from the origin geographical reference area *i* to the destination geographical reference area *j* over a reference period.

In general, an ODM is structured as a table reporting the (a) reference period (timestamp), (b) area of origin, (c) area of destination and (d) count of movements. Such features already yield key aspects to understand the data sets that are not standardized in terms of (a) frequency of the data (i.e. time lag between consecutive samples) and latency (time lag between data collection and availability to the user), (b) granularity of the data at origin and (c) destination, and (d) definition of movement. In addition, it is important to understand how the data are geolocated, whether they refer to information extrapolated to the total population based on the market share, representativity and

[8] More information is available at www.google.com/covid19/mobility/.

[9] More information is available at https://covid19.apple.com/mobility.

the "confidentiality threshold" (i.e. a minimum number of movements under which data are filtered out to decrease de-anonymization risks). All the above introduce diverse levels of uncertainties and heterogeneity, especially in the attempt to make a comparative analysis across data providers and countries. This is further explained by Vespe et al.'s article (2021), where all the aspects above are introduced in the context of MNO data, although most of the considerations can also be extended to other privately held non-traditional data sources.

In particular, data quality aspects are not fully disentangled from privacy, data security and commercial confidentiality. For example, the definition of a confidentiality threshold combined with the selected spatial and temporal resolutions has an impact on the resulting mobility insights as well as on the de-anonymization risk: detailed flows can be mapped between relatively small areas and at high temporal resolution. Nevertheless, the confidentiality threshold would filter out a vast portion of the movements of interest. On the contrary, the effect of the confidentiality threshold on low temporal and spatial resolution data would be lower, at the price of the granularity of the relevant mobility insights. It is therefore important to select such parameters adequately or alternatively understand the limitations of aggregate and anonymized data.

Results: Mobility data products and tools

The first results of the initiative were communicated to the general public by the European Commission daily news[10] on 15 July 2020, by the Joint Research Centre (JRC) news[11] and DG CONNECT.[12]

Mobility data products were developed during the first phase of the initiative and are updated daily. The products are the result of space-time aggregation and normalization of the original ODM data as received by MNOs. This operation allows for comparability of the data set across operators and countries, as well as further desensitization of the insights. The products are currently expanded to feed early warning mechanisms to detect anomalies in usual mobility patterns such as gatherings (De Groeve et al., 2020). Products and tools shared with policymakers and practitioners are described in the following paragraphs.

Mobility indicator

The mobility indicator provides insights into mobility trends in a comparative way – in time and across countries. It results from aggregation and normalization procedures, bringing the data to a common time–frequency (daily) and space granularity (Nomenclature of Territorial Units for Statistics (NUTS)[13]). For each geographical unit, the indicator provides information on mobility disaggregated into inward, outward and internal movements. The indicator, along with data on timing and intensity of the measures in several countries, has been used to assess the impact on mobility of the confinement measures designed to stop the spread of COVID-19.

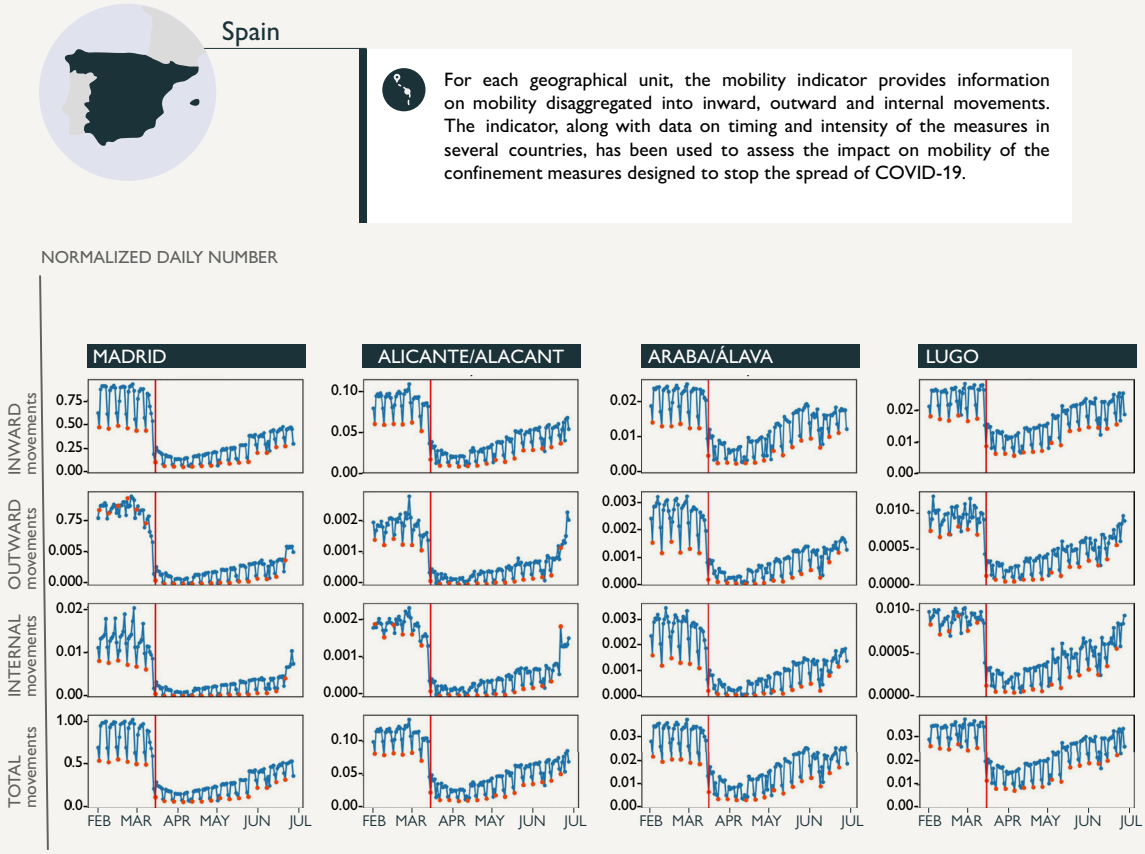
[10] More information is available at https://ec.europa.eu/commission/presscorner/detail/en/mex_20_1359.

[11] More information is available at https://joint-research-centre.ec.europa.eu/jrc-news/coronavirus-mobility-data-provides-insights-virus-spread-and-containment-help-inform-future-2020-07-14_en.

[12] More information is available at https://digital-strategy.ec.europa.eu/en/news/coronavirus-mobility-data-provides-insights-virus-spread-and-containment-help-inform-future.

[13] More information is available at https://ec.europa.eu/eurostat/web/nuts/background.

Figure 1. Mobility indicator (internal – top, inward, outward; and total – bottom) for four NUTS 3 areas in Spain (Madrid – left; Alicante, Araba and Lugo – right)



Source: Santamaria et al., 2020.

Note: The vertical line marks the national lockdown on 15 March 2020, while the red dots indicate Sundays.

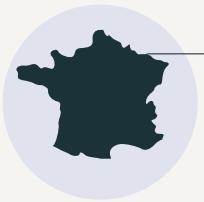
This map is for illustration purposes only. The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by the International Organization for Migration.

The results reported by Santamaria et al. (2020) show that the confinement measures explained up to 90 per cent of the mobility patterns in the reference period. The indicator is also used to compare mobility with the infection reproduction number *Rt*.

Connectivity matrix

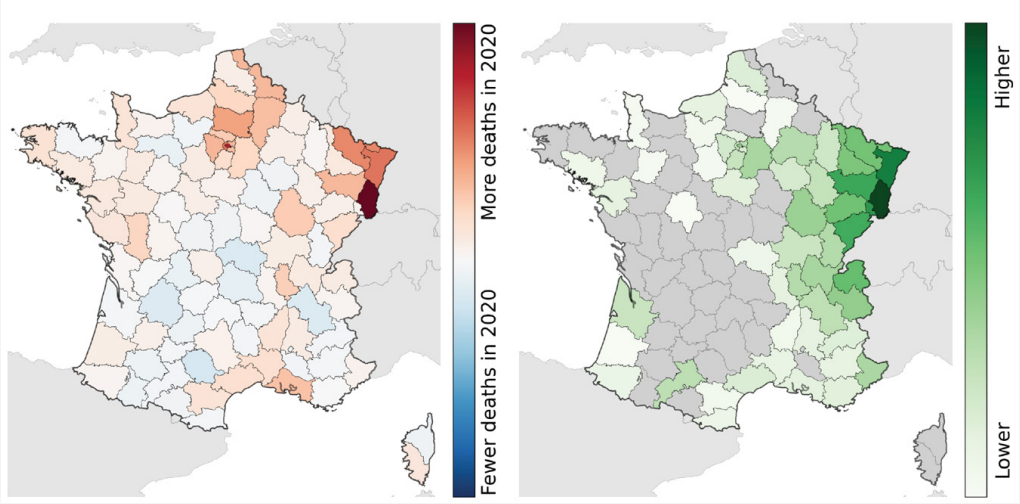
Similar to the mobility indicator, the connectivity matrix aggregates the original ODMs received by MNOs into a common space–time granularity (at NUTS 3 level), but it also provides information about bilateral movements between areas. The connectivity matrix is calculated with a weekly time–frequency (distinguishing between weekday and weekend mobility). As detailed by Iacus et al. (2020a), connectivity matrices have been used to demonstrate that human mobility is highly correlated with the spread of COVID-19 outbreaks through different case studies. In particular, mobility can explain from 52 per cent up to 92 per cent of the excess deaths reasonably linked to the COVID-19 infections in France.

Figure 2. Cumulative number of excess deaths in 2020 versus 2019, from 1 to 25 March (left) and "connectivity" levels from the department of Haut-Rhin (darkest area in the map) during the week of 23–29 February 2020



France

Connectivity matrices were used to demonstrate that human mobility is highly correlated with the spread of COVID-19 outbreaks, through different case studies. In particular, mobility can explain from 52 up to 92 per cent of the excess deaths reasonably linked to the COVID-19 infections in France.



Source: Iacus et al., 2020a.

Note: These maps are for illustration purposes only. The boundaries and names shown and the designations used on these maps do not imply official endorsement or acceptance by the International Organization for Migration.

Similar results were found using connectivity data for Italy and Spain, providing mobility-based evidence that could be useful for scenario-building exercises.

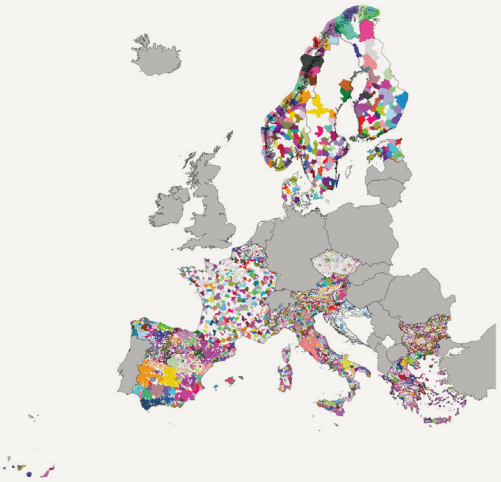
In addition to contributing to predicting future outbreak dynamics, connectivity information between provinces and regions can be used to design targeted response measures in case of future waves of the virus.

Mobility functional areas

MFAs are data-driven geographic zones with a high degree of intermobility exchange and can be input to the definition of tailored measures, ensuring a balance between epidemiological effects and socioeconomic impacts. Mobility patterns shape MFAs in Europe that, in several cases, cross regional or provincial borders. As further explained in the results published by Iacus et al. (2020b, 2021), although slightly changing on a daily basis, MFAs are persistent in time. In addition, confinement measures in Europe have not only reduced the volume of mobility but also changed MFAs.

Figure 3. Mobility functional areas for a preliminary set of countries mapped before confinement measures

Mobility functional areas (MFAs) are data-driven geographic zones with a high degree of intermobility exchange. The mobility patterns shape MFAs in Europe that, in several cases, cross regional or provincial borders. Confinement measures in Europe have not only reduced the volume of mobility but also had a clear impact on the shape of the MFAs.



Source: Iacus et al., 2020b.

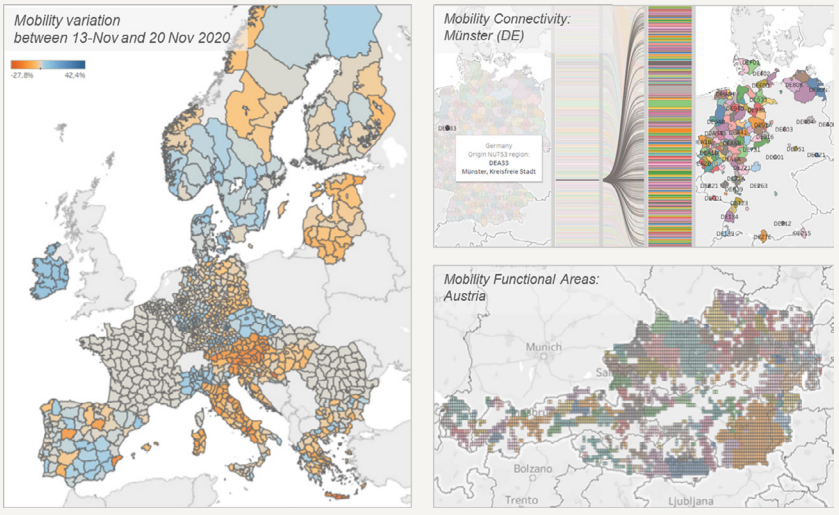
Note: This map is for illustration purposes only. The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by the International Organization for Migration.

Mobility Visualization Platform

Based on these mobility data products, tools have been developed and made available to European Union member States. The Mobility Visualization Platform, in particular, allows almost real-time situational awareness, monitoring the effectiveness of measures on restricting mobility in light of the evolution of the pandemic.

Figure 4. The Mobility Visualization Platform providing access to mobility data products derived from mobile network operator data

The Mobility Visualization Platform has been developed to visualize and analyse mobility trends across regions to inform policy. The Platform presents mobility data products comparable at the national, regional and NUT 3 levels and features European Centre for Disease Prevention and Control (ECDC) data. Access to the Platform is provided to practitioners and policymakers at the European Commission, at ECDC and in European Union member States.



Source: Vespe et al., 2021.

Note: These maps are for illustration purposes only. The boundaries and names shown and the designations used on these maps do not imply official endorsement or acceptance by the International Organization for Migration.

During the COVID-19 emergency situation, the Platform showed operational value also in monitoring the reaction time and impact of targeted lockdowns on mobility at the local, regional and national scale. As mentioned in the strategy on staying safe from COVID-19 during winter, adopted by the European Commission (2020a) in December 2020, "Insights into mobility patterns and role in both the disease spread and containment should ideally feed into such targeted measures. The Commission has used anonymized and aggregated mobile network operators' data to derive mobility insights[14] and build tools to inform better targeted measures, in a Mobility Visualization Platform, available to the Member States. Mobility insights are also useful in monitoring the effectiveness of measures once imposed."

Furthermore, the Platform and the products have been used to provide scenarios and tools for locally targeted COVID-19 non-pharmaceutical intervention measures at the European Union level, as set out in the JRC–ECDC report, which introduces fine-grained targeting measures and tools based on mobility data (De Groeve et al., 2020).

Conclusions

The unique B2G initiative between multiple MNOs and the European Commission outlined several areas worthy of attention when using digital trace data for policy.

Because of the need to promptly react to an emergency situation, the initiative was based mostly on data and products already developed by MNOs. From a data quality perspective, there was a high degree of heterogeneity in the data set (i.e. movements estimated with different algorithms, as well as through a wide range of anonymization processes). This implied substantial scientific analysis to produce comparable insights across countries and operators, ultimately resulting in lessened information content.

Nevertheless, it is important to underline that in this type of initiative, scientific challenges are often only a part of the issues that need to be addressed, which also include aspects of security, commercial sensitivity, fundamental rights, transparency and communication, to name a few.

This unprecedented initiative also demonstrated the importance of an interdisciplinary approach, gathering lawyers, epidemiologists, telecommunication engineers, data scientists, software developers, communication specialists and policymakers. The success of similar initiatives in the future would rely on setting up multidisciplinary groups of experts as well as guaranteeing fast and systematic response to face future crisis situations. With regard to the latter, updated protocols and guidelines need to be already in place at the time of the crisis to avoid potential deadlocks. In light of this, as part of the European Strategy for Data, the Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act) (European Commission, 2020b) aims to foster the availability of data for use by increasing trust in data intermediaries and strengthening data-sharing mechanisms across the European Union. As part of the same strategy, the recent Proposal for a Regulation of the European Parliament and of the Council on harmonized rules

[14] More information is available at https://digital-strategy.ec.europa.eu/en/news/coronavirus-mobility-data-provides-insights-virus-spread-and-containment-help-inform-future.

on fair access to and use of data (Data Act) (European Commission, 2022) aims at maximizing the value of data in the economy by ensuring that a wider range of stakeholders gain control over their data and that more data are available for innovative use.

Finally, efforts should also aim at making insights and derived products publicly available to the research community, in order to harness their full potential and ensure reproducibility of research outcomes.

Acknowledgements

The authors acknowledge the support of European MNOs (among which are 3 Group – part of CK Hutchison, A1 Telekom Austria Group, Altice Portugal, Deutsche Telekom, Orange, Proximus, TIM Telecom Italia, Tele2, Telefonica, Telenor, Telia Company and Vodafone) in providing access to aggregate and anonymized data, an invaluable contribution to the initiative. The authors would also like to acknowledge the GSMA[15] and colleagues from Eurostat[16] and ECDC[17] for their input in drafting the data request.

Finally, the authors would also like to acknowledge the support from JRC colleagues – and in particular the E3 Unit, for setting up a secure environment to host and process the data provided by MNOs, as well as the E6 Unit (the Dynamic Data Hub Team) for their valuable support in setting up the data lake.

[15] GSMA, or the GSM Association, is an organization of mobile network operators.

[16] Eurostat is the statistical office of the European Union.

[17] The European Centre for Disease Prevention and Control is an agency of the European Union.

REFERENCES*

De Groeve, T., A. Annunziato, L. Galbusera, G. Giannopoulos, S. Iacus, M. Vespe, J. Rueda Cantuche, A. Conte, B. Sudre and H. Johnson

2020 *Scenarios and Tools for Locally Targeted COVID-19 Non Pharmaceutical Intervention Measures: Building the Necessary Tools for Monitoring and Planning the Containment of COVID-19 at EU Level*. JRC Science for Policy Report. Publications Office of the European Union, Luxemburg. Available at https://publications.jrc.ec.europa.eu/repository/handle/JRC122800.

European Commission

2020a Communication from the Commission to the European Parliament and the Council: Staying safe from COVID-19 during winter. COM(2020) 786. Available at https://ec.europa.eu/health/sites/health/files/preparedness_response/docs/covid-19_stayingsafe_communication_en.pdf.

2020b Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act). COM(2020) 767. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=COM:2020:767:FIN.

2022 Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act). COM(2022) 65 final. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022PC0068&qid=1647540415418&from=EN.

European Data Protection Board (EDPB)

2020 Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak. Available at https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-042020-use-location-data-and-contact-tracing_en.

GSMA

2021 European Commission and GSMA partners on Data4Covid. 1 February. Available at www.gsma.com/gsmaeurope/resources/d4c/.

Iacus, S.M., C. Santamaria, F. Sermi, S. Spyratos, D. Tarchi and M. Vespe

2020a Human mobility and COVID-19 initial dynamics. *Nonlinear Dynamics*, 101(3):1901–1919.

2020b *Mapping Mobility Functional Areas (MFA) by Using Mobile Positioning Data to Inform COVID-19 Policies: A European Regional Analysis*. JRC Technical Reports. Publications Office of the European Union, Luxembourg. Available at www.preventionweb.net/publication/mapping-mobility-functional-areas-mfa-using-mobile-positioning-data-inform-covid-19.

2021 Mobility functional areas and COVID-19 spread. *Transportation*.

** All hyperlinks were working at the time of writing this report.*

Mamei, M., N. Bicocchi, M. Lippi, S. Mariani and F. Zambonelli

2019 Evaluating origin–destination matrices obtained from CDR data. *Sensors*, 19(20):4470.

Rango, M. and M. Vespe.

2017 Big Data and alternative data sources on migration: From case-studies to policy support. Summary report. Ispra.

2021 Data innovation for migration: Why now and how? OECD Development Matters, 23 January. Available at https://oecd-development-matters.org/2021/01/23/data-innovation-for-migration-why-now-and-how/.

Santamaria, C., F. Sermi, S. Spyratos, S.M. Iacus, A. Annunziato, D. Tarchi and M. Vespe

2020 Measuring the impact of COVID-19 confinement measures on human mobility using mobile positioning data. A European regional analysis. *Safety Science*, 132:104925.

Spyratos, S., M. Vespe, F. Natale, I. Weber, E. Zagheni and M. Rango

2018 Migration data using social media: A European perspective. JRC Science Hub.

2019 Quantifying international human mobility patterns using Facebook Network data. *PLOS ONE*, 14(10):e0224134.

Vespe, M., S.M. Iacus, C. Santamaria, F. Sermi and S. Spyatos

2021 On the use of data from multiple mobile network operators in Europe to fight COVID-19. *Data & Policy*, 3(9).

16

MOBILE PHONES

SOCIAL MEDIA

ARTIFICIAL INTELLIGENCE/ MACHINE LEARNING

SATELLITE DATA

OTHER

BUILDING DATA PARTNERSHIPS

OPERATIONALIZING DATA COLLABORATIVES FOR MIGRATION

Stefaan G. Verhulst[1] *and Andrew Young*[2]

Today's migration data and information challenges

Today, policymakers, international organizations and civil society actors working in the field of migration face a wide array of critical information challenges, limiting their mission to make decisions and provide services in the best interests of migrants, especially those in vulnerable situations. Some of these challenges are included in the following non-exhaustive list:[3]

(a) The **complexity and number of variables involved**, and how they interact in difficult and hard-to-predict ways. While existing platforms, like the Migration Data Portal,[4] seek to provide a one-stop shop for a wide variety of data points on migration, useful data for the many variables associated with it can be lacking or inaccessible.

(b) The **institutional and official statistical capacity** of several countries are not consistent, making it difficult for institutions to make evidence-based decisions about migration (European Commission, n.d.).

(c) Migration-related data are often **not disaggregated by important demographic and socioeconomic variables** like sex, age or profession (Nielsen, 2014).

(d) Data that can inform migration policy and service delivery are often **collected and stored in a fragmented manner**, raising issues related to interoperability and limiting the ability of policymakers to pull together the broad, diverse information necessary for making migration-related decisions (Franklinos et al., 2021).

[1] Stefaan G. Verhulst is Co-Founder and Chief of Research and Development Officer of the Governance Lab (The GovLab). He is Editor-in-Chief of the journal *Data & Policy* (Cambridge University Press), Senior Advisor to the Markle Foundation, and a partner in the Big Data for Migration Alliance. He is also an Adjunct Professor in the Department of Media, Culture and Communications at New York University, a Senior Research Fellow at Central European University, and an Affiliated Senior Research Fellow at the University of Pennsylvania.

[2] Andrew Young is Global Technology Industry Analyst at Ernst & Young. Previously he worked as Knowledge Director of the GovLab at New York University, where he also received his master's degree in Media, Culture and Communication, focusing on Technology and Society.

[3] See, for instance, the comparative overview at the Migration Data Portal of the different types of migration data and the current gaps and limitations, available at https://migrationdataportal.org/themes/migration-data-sources.

[4] More information is available at https://migrationdataportal.org.

Data collaboratives as a means to address some of today's data challenges

While the information challenges associated with migration are real, and limit the ability of policymakers and civil society to make effective, targeted responses, recent years have seen growing efforts on the part of governments, aid agencies and others to collect and make better use of data. It is, however, when combined with other sources of data – often collected by the private sector – that these efforts become most effective. We call such partnerships *data collaboratives*.

We use the term "data collaboratives" to refer to an emergent form of public–private partnership that allows for collaboration and access to data for reuse across sectors and actors. This model has now been used in a variety of sectors and geographies, ranging from accessing environment-related data to accelerate disease treatments (The GovLab, n.d.a) to leveraging private bus data to improve urban planning (The GovLab, n.d.b). When designed responsibly, data collaboratives have the potential to help overcome the problem of data fragmentation and fill gaps in existing information architectures for migration, allowing policymakers and civil society organizations to drive more targeted support and policies and generate new knowledge on migrant populations across the globe. In particular, as it relates to migration, data collaboratives have the potential to:

(a) **Improve situational analysis related to migrant and refugee movements**. In various countries around the world, satellite imagery and social media data, along with other forms of data, are being used to improve institutional awareness and response to shifting migration patterns (Iacus, 2017). For instance, the Governance Lab, together with the United Nations Children's Fund (UNICEF) and other partners, with support from the World Bank (2018), established a data collaborative focused on internally displaced people in Somalia. The initiative leveraged satellite imagery data and other private-sector data sets to increase our understanding of where and how displaced people – and children in particular – moved about the country.

(b) **Generate new knowledge on root causes and drivers of migration and enable the transfer of existing knowledge across sectors**. Researchers around the world have increasingly used web data, including social media data and search query data, to gain new insights that can inform more evidence-based policymaking and responsive service delivery. The Data Challenge on Integration of Migrants in Cities, for example, was a European Commission (Joint Research Centre) initiative seeking insight into the integration and concentration of migrants in cities.[5] It provided access to high spatial resolution census data that show the concentration of migrants in cities across eight European Union States, and encouraged participating teams to examine patterns and develop new insights that can aid in the integration and understanding of migrant communities across the European Union.

(c) **Inform prediction and forecasting related to migration and refugees**. A June 2017 study from the Pew Research Center showed how different search patterns in certain regions – e.g. Arabic-language searches arising from Türkiye including the keyword "Greece" – correlated with changes in migration flows to Europe (Connor, 2017). A research group at the Office of the United Nations High Commissioner for Refugees used weather data to forecast migration

[5] More information on Data for Integration (D4I) is available at https://bluehub.jrc.ec.europa.eu/datachallenge/.

flows, coordinating with partners at national migration agencies and meteorological offices to anticipate how weather shifts could impact cross-border movement and needs at refugee camps (Migration Data Portal, 2021a). A more systematized (and responsible) approach to cross-sector data-sharing and analysis, including but not limited to search query data, could provide these types of predictive insights and inform more anticipatory and effective responses to shifts in migration patterns.

(d) **Enable more targeted impact assessment and evaluation of migration interventions and responses**. Several data collaboratives have been launched in recent years to better understand the impact and effectiveness of migration-related policies and interventions. The SoBigData exploratory migration studies project, funded by the European Union's Horizon 2020 research and innovation programme, for example, was leveraging data collaboration to answer questions related to evaluating migration policy in Europe and migration more generally.[6] In particular, in light of COVID-19, several data collaboration efforts used location data to understand internal migration flows and mobility patterns in response to social distancing and lockdown policies. The COVID-19 Mobility Monitoring project, for example, used aggregated location data from mobile phones to show the fluctuations in traffic flows between national provinces.[7]

Different types of data collaboratives for different purposes

Our research has informed an emerging typology of data collaboratives, comprising six core operational models that practitioners could use to match the supply of data to the demand for it in the migration space (Verhulst et al., 2019):

(a) *Public interfaces* – Data holders, usually a single company, provide open access to certain data assets, including preprocessed data and data-driven tools like maps or dashboards. This type of data collaborative enables independent uses of the data by external parties. Public interfaces are usually formed with a target audience or type of use case, though the data assets are made publicly accessible. The two most common approaches to public interfaces are application programming interfaces (APIs) and data platforms. APIs are digital protocols that publish data in an automated fashion on a near real-time data basis, allowing for open, ongoing access to data independent of the direct involvement of the data provider. Climate LLC's FieldView products[8] and Reddit's API for public health research (Park and Conway, 2018) are two examples of APIs in the field, along with data platforms like Open Diversity Data and Waze Connected Citizens. Data platforms, like Double Union's Open Diversity Data[9] or Uber Movement,[10] enable public access to private-sector data assets and tools. While data platforms and APIs have equally flexible uses, data platforms often require less data or software development expertise to set up.

(b) *Data pooling* – In this practice area, data holders agree to create a unified presentation of data sets as a collection accessible by multiple parties. This system can promote highly cooperative

[6] More information is available at www.sobigdata.eu/exploratories/migration-studies.

[7] More information is available at https://covid19mm.github.io/.

[8] More information is available at www.climate.com/.

[9] More information is available at http://opendiversitydata.org.

[10] More information is available at https://movement.uber.com/?lang=en-US.

or more independent work depending on the levels of access and the objectives of the data pool. Data pools can be either public or private in nature. Public data pools draw together assets from multiple pre-approved data holders across the public and private sectors, and make these shared assets available to all via the Internet. Although public data pools are more often than not open access, they are developed specifically with certain audiences or intended use cases. Examples of public data pools include Accelerating Medicines Partnership,[11] Global Fishing Watch,[12] Global Forest Watch[13] and the Humanitarian Data Exchange.[14] Private data pools, unlike public data pools, pool assets in a controlled and restricted access environment. This limits data contribution and data access exclusively to approved partners. Private data pools, such as Mobile Data, Environmental Extremes and Population (MDEEP)[15] and OpenTraffic,[16] are highly topic-specific and therefore aimed towards particular user groups.

(c) *Prizes and challenges* – Data holders make data available to participants who compete to develop apps, answer problem statements, test hypotheses and premises, or pioneer innovative uses of data for the public interest and to provide business value. The intention of the prizes and challenges is to use and target expertise to address challenges or opportunities defined by the project organizer (Verhulst, 2015). There are two main kinds of innovation challenges: open and selective. In an open innovation challenge, organizers give participants open access to data sets to lower barriers to entry in terms of data access and analysis. While open innovation challenges are directed towards a particular issue, the more independent nature of the challenge opens up the risk for projects falling beyond the intended scope. In 2017, for example, LinkedIn hosted the Economic Graph Challenge, wherein applicants submitted a proposal to the company for use of their data. Winning proposals were given access to LinkedIn's data for six months, as well as USD 25,000 as research prize.[17] In selective innovation challenges, participants are given restricted access to data only after their proposed projects are approved for the challenge. This means that the multiple parties involved collaborate more closely over the course of the challenge, and the scope of data available to the participants can vary depending on their projects. Recent selective innovation challenges include Data for Climate Action,[18] GBDX for Sustainability Challenge,[19] Türk Telekom Data for Refugees (D4R) Challenge[20] and Orange Telecom Data for Development (D4D) Challenge.[21]

(d) *Trusted intermediary* – Trusted third-party intermediaries bridge the gap between data holders and data users working in the public interest by granting data users access to sensitive data under strict access controls. Intermediaries can play a role in facilitating data collaborations by matching supply and demand actors. Some go further to provide technical expertise in the form of data analysis to achieve actionable insights. Data brokerages and third-party analytics projects are two models for trusted intermediaries. As the name suggests, data

[11] More information is available at https://datacollaboratives.org/cases/accelerating-medicines-partnership-amp.html.
[12] More information is available at https://datacollaboratives.org/cases/global-fishing-watch.html.
[13] More information is available at https://datacollaboratives.org/cases/global-forest-watch.html.
[14] More information is available at https://datacollaboratives.org/cases/the-humanitarian-data-exchange-hdx.html.
[15] More information is available at https://datacollaboratives.org/cases/mobile-data-environmental-extremes-and-population-mdeep-project.html.
[16] More information is available at https://icos.urenio.org/applications/opentraffic/.
[17] More information is available at https://specialedition.linkedin.com/details.
[18] More information is available at www.unglobalpulse.org/data-for-climate-action.
[19] More information is available at https://datacollaboratives.org/cases/gbdx-for-sustainability-challenge.html.
[20] More information is available at https://datacollaboratives.org/cases/turk-telekom-data-for-refugees-d4r-challenge.html.
[21] More information is available at https://datacollaboratives.org/cases/orange-telecom-data-for-development-challenge.html.

brokerages use third-party brokers to establish collaborations and to match supply-side actors with demand-side clients. These collaborations are generally time and purpose bound. The Consumer Data Research Centre in the United Kingdom acts as a trusted intermediary to enable access for researchers to data held by consumer-related businesses.[22] In third-party analytics projects, on the other hand, trusted intermediaries access private-sector data to conduct targeted analyses and share only the insights derived from their study with the public. These projects generally work on highly sensitive personal data (i.e. call detail records). An example of a third-party analytics project would be Dalberg's (2018) Food Security Manager, which analyses privately held big data to help development and humanitarian organizations make data-driven programme decisions.

(e) *Research and analysis partnerships* – Here data holders engage directly with public-sector partners and share certain proprietary data assets to generate new knowledge with public value. This is a highly cooperative practice used by businesses to augment existing business capabilities, incubate new product ideas, or analyse questions outside the scope of their internal business operations. Research and analysis partnerships take on many forms depending on the actors involved, though the two more common partnership models are data transfers and data fellowships. In a data transfer, the company provides data for analysis in a highly restrictive system. Often presented as a sort of "data philanthropy", data transfers have specific, approved uses that are decided through agreements between the involved parties. An example of a data transfer is Cuebiq's Data for Good's partnership with MIT Media Lab.[23] Cuebiq, a location intelligence company, shared its location data with researchers at MIT Media Lab for use in their Atlas of Inequality project. Data fellowships, on the other hand, create opportunities for individuals or parties to access data assets at companies, or provide data science expertise to the partner institution for a fixed period of time. Amazon Web Services (AWS) and Azavea Open Source, for example, operate a joint fellowship programme.[24] Azavea, a social enterprise, works with AWS to recruit fellows to work on various civic, social and economic projects using geospatial technology. Fellows are given access to AWS Earth data, such as Landsat and Next Generation Weather Radar (NEXRAD) data, to name a few.

(f) *Intelligence generation* – In the intelligence generation model, data holders internally develop data-driven analyses, tools, and other resources, and release those insights to the broader public. No data is shared with external parties. Instead, internal actors work on highly directed projects to develop insights and analyses to support policymaking and service-delivery efforts. Although there is no direct cross-sector sharing, intelligence generation allows for knowledge transfer and demonstrates the potential public value of the analysis of private-sector data assets. The JPMorgan Chase Institute's insight reports are an example of intelligence generation, wherein a private-sector actor is using its expertise and proprietary data from the multinational financial corporation to produce research reports on economic issues.[25] Their work is aimed at helping policymakers, businesses and non-profit organizations address economic and financial issues.

[22] More information is available at www.cdrc.ac.uk/.

[23] More information is available at www.cuebiq.com/press/cuebiq-launches-data-good-initiative-collaborates-mit-20-universities-nonprofits-worldwide/.

[24] More information is available at https://fellowship.azavea.com/about/.

[25] More information is available at www.jpmorganchase.com/institute.

Four steps towards establishing data collaboratives for migration

The public value of cross-sector data collaboration in the realm of migration is starting to become apparent in diverse and ongoing cross-sector data-sharing experiments. Needless to say, much remains to be done to make these efforts more systematic. As it stands, decision makers on both the supply and demand sides of data collaboratives often lack a clear and actionable understanding of whether to establish data collaboratives, when and how – and, importantly, how to do so in a responsible manner that does not create privacy and other risks. Without a clear understanding of both the risks and rewards of accessing data across sectors, data holders are likely to remain risk averse and restrict the flow of data, thus minimizing the positive secondary usage and societal impacts of data.

To help practitioners operationalize data collaboratives to improve outcomes in the field of migration, we suggest below four necessary steps to transition from a series of innovative yet ad hoc projects to a more systematic framework of action. Together, these steps provide the foundations that could allow all stakeholders – in government, civil society and the private sector – to design effective, more targeted policy interventions.

(a) First, **prioritize purpose-driven data collaboration for migration**, defining and acting upon the most impactful questions related to migration. Data collaboratives, across sectors, often suffer from a poor articulation of specific problems, needs, and data gaps that public-interest actors face and could be addressed through data collaboration. More targeted questions related to migration can inform the design of data collaboration efforts that will responsibly and ethically use new data sources to inform migration policies and programmes around the world. They can also help data holders across sectors better understand when and how their data could make a difference. The GovLab's 100 Questions Initiative seeks to do just this, developing a process that takes advantage of a range of expertise on given topics or domains so as to identify and prioritize those questions that are high impact, novel and feasible. Early in 2020, 10 questions related to migration were modified and selected by "bilinguals", or individuals who are proficient in a given domain and also skilled in the fields of data science and/or statistics, such as "What are the current trends in human mobility globally? How many people migrate every year? What policies, evidence and programmes (can) depolarize public debate or reduce anti-immigrant sentiment?"[26] By encouraging this practice, data-driven projects will arise out of genuine need, matching the supply of data to real demand.

(b) Second, **empower and nurture "data stewards"**. As is the case with data collaboration in general, some leading innovators are already establishing concrete roles, responsibilities and practices for determining whether to share data for the public good, when and how. Organizations across sectors can work to professionalize the supply side of migration-focused data collaboratives and increase the resilience of such efforts by establishing data stewards – either individual leaders or teams – to put this work into practice. Our research, and the facilitation of a global Data Stewards Network, has shown that data stewards often take on five key roles in their institutions: (a) partnership and community engagement; (b) internal

[26] More information is available at https://migration.the100questions.org/.

coordination and staff engagement; (c) data audit, ethics, and assessment of value and risk; (d) dissemination and communication of findings; and (e) nurturing data collaboratives to sustainability (Verhulst et al., 2020).

(c) Third, **determine fit-for-purpose operational and governance models for data collaboration**. There is no one-size-fits-all approach to data collaboration. These initiatives can take many forms, and successful efforts are dependent on the specific needs, opportunities, and constraints of actors on the supply and demand sides of the collaboration. Within the context of our work with the Big Data for Migration (BD4M) Alliance, we have developed a distinctive methodology (studios) to design data collaboratives that are fit for purpose. Our framework seeks to identify the potential incentives, challenges, and enabling conditions for both data users and data holders. Through distinctive "sprints", we convene a curated group of stakeholders to assess the following:

(i) Problems to be addressed: What are the key questions? What is the context?

(ii) Data sets: What are the minimum viable data sets needed? What are the required characteristics?

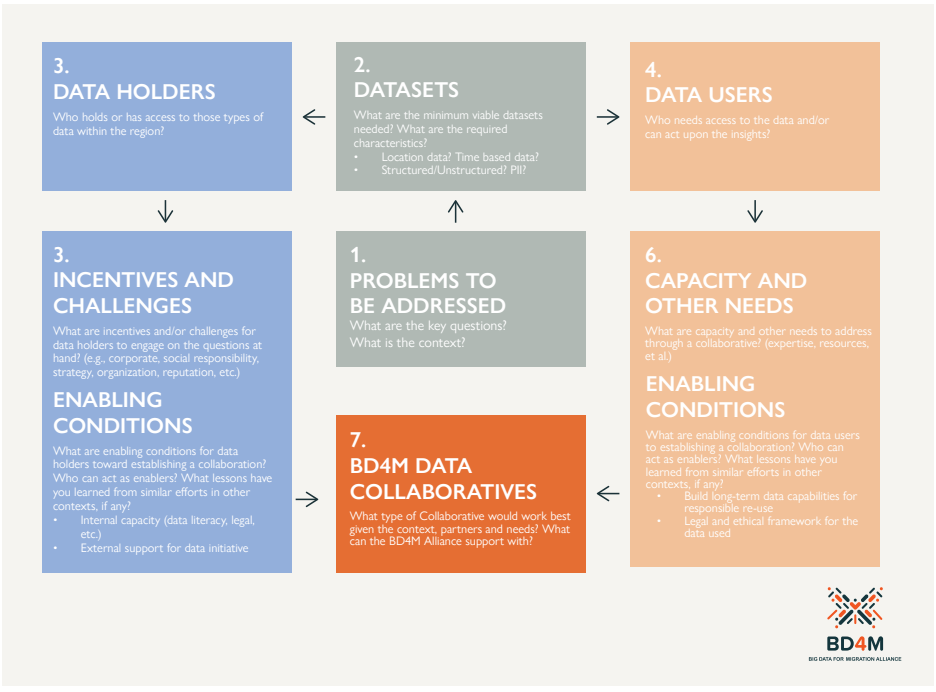
(iii) Data holders: Who holds or has access to those types of data within the region?

(iv) Data users: Who needs access to the data and/or can act upon the insights?

(v) Incentives and challenges and enabling conditions (for data holders): What are the incentives and/or challenges for data holders to engage with the questions at hand? What are the enabling conditions for data holders towards establishing a collaboration? Who can act as enablers? What lessons have you learned from similar efforts in other contexts, if any?

(vi) Capacity and other needs and enabling conditions (for data users): What are the capacities and/or needs to address through a collaborative? What are the enabling conditions for data users to establish data collaboration? Who can act as enablers? What lessons have you learned from similar efforts in other contexts, if any?

(vii) Potential BD4M data collaboratives: What type of collaborative would work best given the context, partners and needs? What can the BD4M Alliance support with?



(d) Finally, **develop data responsibility frameworks for collaborating and sharing data**. As described above, data collaboratives, especially those involving vulnerable communities, are not free of risk. The relatively limited use to date and the lack of systematic data-sharing by corporations and other data holders are likely, at least in part, based on recognition of these risks. To help both the supply- and demand-side actors of data collaboratives more effectively navigate these risks, the research, policy, and technology communities should develop new methodologies, tools, and frameworks to enable the responsible and systematic sharing of data. In the past year, there has been significant progress made within the data ecosystem to define such frameworks. In 2019, the United Nations Office for the Coordination of Humanitarian Affairs released a set of Data Responsibility Guidelines to address the challenges of managing risk in the sharing of personal and sensitive data (Campo, 2019). Responsible Data for Children (RD4C), a joint effort between UNICEF and the GovLab, developed a set of principles, insights and approaches to encourage responsible data management for children.[27] The GovLab also released a Data Responsibility Journey mapping tool in 2021 to help practitioners assess the opportunities and risks to consider at each stage of the data life cycle when implementing a data collaborative.[28] Building upon emerging data responsibility frameworks (Raymond et al., 2016; Berens et al., 2016), operational guidance for data responsibility could decrease the transaction costs, time and energy currently needed to establish data collaboratives between the private and public sectors, and do so in a way that does not create undue risks to the intended beneficiaries of these public–private data-sharing arrangements. Important to recognize, however, is that the governance model or framework used to guide decision-making in a data collaborative is highly context dependent. Data stewards across sectors are experimenting with several emerging governance models for data reuse (such as ethical councils and independent review boards), as well as more traditional approaches (such as contracts and terms and conditions).

[27] More information is available at https://rd4c.org/.

[28] More information is available at https://dataresponsibilityjourney.org.

REFERENCES*

Berens, J., U. Mans and S. Verhulst

2016 *Mapping and Comparing Responsible Data Approaches*. The GovLab and Centre for Innovation, Leiden University. Available at www.thegovlab.org/static/files/publications/ocha.pdf.

Campo, S.

2019 Introducing the Working Draft of the OCHA Data Responsibility Guidelines. Centre for Humanitarian Data, 7 March. Available at https://centre.humdata.org/introducing-the-working-draft-of-the-ocha-data-responsibility-guidelines/.

Connor, P.

2017 The digital footprint of Europe's refugees. Pew Research Center, 8 June. Available at www.pewglobal.org/2017/06/08/digital-footprint-of-europes-refugees/.

Dalberg

2018 Dalberg Data Insights identifies areas at-risk for food insecurity using mobile phone data. 22 June. Available at https://dalberg.com/our-ideas/dalberg-data-insights-identifies-areas-risk-food-insecurity-using-mobile-phone-data/.

European Commission

n.d. EURODAC. Available at https://home-affairs.ec.europa.eu/pages/glossary/eurodac_en.

Franklinos, L., R. Parrish, R. Burns, A. Caflisch, B. Mallick, T. Rahman, V. Routsis, A. Sebastián López, A. Tatem and R. Trigwell

2021 Key opportunities and challenges for the use of big data in migration research and policy. *UCL Open: Environment Preprint*.

Iacus, S.M.

2017 To which extent social media can help migration monitoring? Measuring Migration: NTTS 2017 satellite event, 13 March. Available at https://ec.europa.eu/eurostat/cros/ntts2017programme/data/abstracts/abstract_221.html.

Migration Data Portal

2021 UNHCR's "Winter Cell": Forecasting migration flows with weather data. 27 May. Available at https://migrationdataportal.org/data-innovation/unhcrs-winter-cell-forecasting-migration-flows-weather-data-0.

Nielsen, R.C.

2014 Big data and international migration. United Nations Global Pulse, 16 June. Available at www.unglobalpulse.org/big-data-migration.

** All hyperlinks were working at the time of writing this report.*

Park, A. and M. Conway

2018 Tracking health related discussions on Reddit for public health applications. *AMIA Annual Symposium Proceedings*, 2017(April):1362–1371.

Raymond, N., Z. Al Achkar, S. Verhulst and J. Berens

2016 *Building Data Responsibility into Humanitarian Action*. United Nations Office for the Coordination of Humanitarian Affairs (OCHA) Policy and Study Series, May. Available at www.unocha.org/publication/policy-briefs-studies/building-data-responsibility-humanitarian-action.

The Governance Lab (The GovLab)

n.d.a Accelerating Medicines Partnership (AMP). Data Collaboratives. Available at http://datacollaboratives.org/cases/accelerating-medicines-partnership-amp.html.

n.d.b Beeline Crowdsourced Bus Service. Data Collaboratives. Available at http://datacollaboratives.org/cases/beeline-crowdsourced-bus-service.html.

Verhulst, S.G.

2015 Governing through prizes and challenges. Medium, 28 January. Available at https://medium.com/@sverhulst/governing-through-prizes-and-challenges-677f3ef861d1.

Verhulst, S.G., A. Young, M. Winowatan and A.J. Zahuranec

2019 *Leveraging Private Data for Public Good: A Descriptive Analysis and Typology of Existing Practices*. The GovLab, October. Available at https://datacollaboratives.org/existing-practices.html.

Verhulst, S.G., A.J. Zahuranec, A. Young and M. Winowatan

2020 Wanted: Data stewards – (Re-)defining the roles and responsibilities of data stewards for an age of data collaboration. The GovLab, March. Available at www.thegovlab.org/static/files/publications/wanted-data-stewards.pdf.

World Bank

2018 Announcing funding for 12 development data innovation projects. 29 January. Available at http://blogs.worldbank.org/opendata/announcing-funding-12-development-data-innovation-projects?CID=DEC_TT_data_EN_EXT.

ADDITIONAL READING

DigitalGlobe

2013 DigitalGlobe imagery assists UNHRC in tracking Sudanese refugees. Available at https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/65/DG-REFMAP-CS_WEB_0.pdf.

Maxmen, M.
2017 Out of the Syrian crisis, a data revolution takes shape. *Nature*, 25 October. Available at www.scientificamerican.com/article/out-of-the-syrian-crisis-a-data-revolution-takes-shape/.

Migration Data Portal
2021 Social interactions in Mexico during the swine flu pandemic: Using mobile phone data to measure the effectiveness of human mobility regulations for reducing the virus spread. 26 May. Available at https://migrationdataportal.org/data-innovation-1.

Nature
2017 Data on movements of refugees and migrants are flawed. 1 March. Available at www.nature.com/news/data-on-movements-of-refugees-and-migrants-are-flawed-1.21568.

Rampazzo, F. and I. Weber
2020 Facebook advertising data in Africa. In: *Migration in West and North Africa and across the Mediterranean: Trends, Risks, Development and Governance* (Fargues, P., M. Rango, E. Börgnas and I. Schöfberger, eds.). IOM, Geneva, pp. 32–40. Available at https://publications.iom.int/books/migration-west-and-north-africa-and-across-mediterranean.

Rango, M. and M. Vespe
2017 Big data and alternative data sources on migration: From case studies to policy support. Available at www.researchgate.net/publication/323129786_Big_Data_and_alternative_data_sources_on_migration_from_case-studies_to_policy_support_-_Summary_report.

Song, T.M., J. Song, J.Y. An, L.L. Hayman and J.M. Woo
2013 Psychological and social factors affecting Internet searches on suicide in Korea: A big data analysis of Google search trends. *Yonsei Medical Journal*, 55(1):254–263.

United Nations Children's Fund (UNICEF)
2013 *Tracking Anti-Vaccination Sentiment in Eastern European Social Media Networks*. Regional Office for Central and Eastern Europe and the Commonwealth of Independent States, April. Available at www.unicef.org/eca/reports/tracking-anti-vaccination-sentiment-eastern-european-social-media-networks.

University of Southampton
2016 New global migration mapping to help fight against infectious diseases. *ScienceDaily*, 22 August. Available at www.sciencedaily.com/releases/2016/08/160822083620.htm.

Verhulst, S.G. and A. Young.
2017 *The Potential of Social Media Intelligence to Improve People's Lives: Social Media Data for Good*. The GovLab, New York University. Available at http://datacollaboratives.org/social-media.html.

IOM
UN MIGRATION

International Organization for Migration
17 route des Morillons, P.O. Box 17, 1211 Geneva 19, Switzerland
Tel.: +41 22 717 9111 • Fax: +41 22 798 6150
Email: hq@iom.int • Website: www.iom.int