

ORIGINAL ARTICLE

Open Access



Measuring migration 2.0: a review of digital data sources

Jasper Tjaden^{1,2*} 

*Correspondence:
jasper.tjaden@uni-potsdam.de

¹ Department of Economics
and Social Sciences,
University of Potsdam,
August-Bebel-Str.89,
14482 Potsdam, Germany
Full list of author information
is available at the end of the
article

Abstract

The interest in human migration is at its all-time high, yet data to measure migration is notoriously limited. “Big data” or “digital trace data” have emerged as new sources of migration measurement complementing ‘traditional’ census, administrative and survey data. This paper reviews the strengths and weaknesses of eight novel, digital data sources along five domains: reliability, validity, scope, access and ethics. The review highlights the opportunities for migration scholars but also stresses the ethical and empirical challenges. This review intends to be of service to researchers and policy analysts alike and help them navigate this new and increasingly complex field.

Keywords: Migration, Big data, Digital trace, Measurement, Survey, Review

Introduction

International interest in measuring human migration is at an all-time high. The number of people living in a country other than their country of birth reached an estimated 272 million in 2019, an increase of 51 million from 2010 (UN DESA, 2019). The number of forcibly displaced people due to conflicts and disasters is at its historic high (UNHCR 2020). International migration is expected to continue increasing given higher levels of interconnectedness in the world due to improved communication and transport systems, protracted crises producing displacement, and structural changes such as climate change and population growth in certain world regions.

Governments worldwide have a keen interest in anticipating future migration flows and understanding the drivers of migration to plan ahead, allocate funds, attract workers and students, use remittances, facilitate migrant integration, and manage public opinion, among other issues. The increased demand for systematic measurement from policy-makers has also manifested itself in two landmark policy frameworks adopted in the last

decade: The Global Compact on Migration (GCM) and the Sustainable Development Goals.¹

Second, in step with the increased salience of migration in policy circles, migration research output has grown dramatically (Pisarevskaya et al., 2019). Between 1960 and 1980, the number of academic journals on related subjects quadrupled. In 2020, the International Organization for Migration (IOM) identified over 130 migration-related journals publishing more than 2000 journal articles in English, French or Spanish (IOM, 2020).

Policy and scholarly interest both rely on fundamental measurements of migration, i.e. how many people migrate (flows) or have migrated (stocks) within a specific time frame. Yet, the popularity and relevance of migration has outpaced substantial improvements in the systematic measurement of migration, especially at the global level. Indeed, the demand for ‘evidence’ has revived long-standing calls for better data on international migration which experts have been lamenting for decades (Bilsborrow et al., 1997; Clemens et al., 2009; Laczko, 2016; Lemaitre, 2005; Willekens et al., 2017).²

This is the context in which a set of new data sources emerged providing migration researchers of all disciplines with new opportunities to measure migration. The arrival of “innovative data sources”—often referred to as “Big Data” or “digital trace data”—have been described as a “migration data revolution” (Laczko & Rango, 2014) and bears much potential to complement traditional migration data (Cesare et al., 2018; Sirbu et al., 2021). At the same time, digital data present a host of new ethical challenges for researchers that are of great concern (Beduschi, 2020; Brayne, 2018; Hayes, 2017; Latonero & Kift, 2018; Leese et al., 2021; Molnar, 2019; Zwitter, 2014).

As new researchers and policymakers flock to the field of migration and the empirical study of migration diversifies, there is a need to explain and review new migration data sources to provide a better understanding of their respective limitations and strengths. This paper reviews eight data sources in terms of their reliability, validity, scope for research, access and ethics. The aim is to familiarize experienced and incoming migration scholars with new approaches. The review should be considered an attempt to contribute towards a broader process of interdisciplinary dialogue and expanding the empirical toolbox in migration studies.

Analytical framework

Defining migration

Before mapping novel data sources, it is important to first define migration for the purpose of this study and clearly delineate the scope of the review.

¹ The GCM is the first comprehensive, inter-governmentally negotiated (non-binding) treaty on migration adopted in 2018 by 152 countries, highlights the importance of data and evidence throughout its 23 objectives. The first goal of the GCM itself is improving migration data across the board. The SDGs are the follow-up to the UN Millennium Goals and feature several migration-relevant goals and include indicators to measure progress towards them. New approaches to measuring migration based on “innovative” data sources have mainly been pioneered by academic scholars, yet some of this work has long caught the attention of key national and international governmental stakeholders. In 2009, the United Nations Global Pulse initiative was launched to use new data sources for analysis of development projects and processes, including migration issues. In 2018, the European Commission, in partnership with the IOM Global Migration Data Analysis Centre, launched the ‘Big Data for Migration Alliance’ with the aim of sharing knowledge on data innovation in the field of migration, providing technical support to local and national administrations interested in using new data sources, and testing new data applications for specific policy needs.

² There are exceptions. The collection, harmonisation and estimation of migration data within the EU (Poulain et al., 2006; Raymer et al., 2013) and compilation of data on migration flows in Latin America (Lemaitre, 2005) are key achievements. Data on migration in many non-OECD countries remains limited.

Important aspects of definitions of migration are space (internal or domestic vs. international/ cross-border migration), time (short term vs. long term), type (e.g. labour, irregular, forced, family, education) and form (flows vs. stocks) (Bilsborrow, 2016; UN DESA, 2017). This review takes a broad and inclusive view in line with its aim to describe a menu of novel data sources for diverse groups of migration scholars and research interests. Here, migration will be defined as the changing of residence of an individual within or outside the boundaries of a country for longer than three months. While this definition is broad, it excludes certain types of mobility such as travel for the purpose of recreation, holiday, visits to friends or relatives, business, medical treatment or religious pilgrimage (which usually do not imply a change of residence).³ The review considers data sources providing information on both migration flows (i.e. the number of migrants entering and leaving (inflow and outflow) a country over the course of a specific period, for example, one year) and migration stocks (i.e. the total number of migrants present in a given location at a particular point in time) (see Global Migration Group, 2017). The review considers any form and channel of migration including, among others, labour, family, forced and irregular migration and is not restricted geographically. In addition, to actual observed migration, the review also considers proxies for migration; such as, migration intentions, plans, desires, or aspirations that are commonly used to predict a future change of residence (Tjaden et al., 2019).⁴ The review does not include how novel data can be used to research other fields of interest to migration scholars such as integration, the causes of migration, communication, or the impact of migration on society. These fields are not primarily concerned with or rely on (directly or indirectly) inferring migration flows or stocks from data.

Defining “novel” data sources

What are these “new” data sources and what makes them “newer” than the “old” sources? Two popular concepts are helpful to delineate traditional from “innovative” data sources for migration: “big data” and “digital trace data”.

Big data is commonly defined by the “three V’s”: volume, velocity, variety. Volume refers to the magnitude of data. However, there is “little consensus around the fundamental question of how big the data has to be to qualify as “big data” (Gandomi & Haider, 2015: 137). Velocity refers to the rate at which data are generated which has dramatically increased with the proliferation of digital devices such as smartphones and sensors. Variety refers to the type of data that is being generated. Big data often includes numerical data, text data, images, audio and geo-location data. These Vs are useful to describe many of the sources that are commonly associated with big data such as social media platforms (Facebook, Twitter, Twitter, Instagram etc.) and online search platforms (google). These data compile millions of records about their users ranging from location, online activity to demographic user profile information. At the same time, this information is accessible in real-time, sometimes even publicly through application

³ Mobility (migration regardless of distance and borders) more broadly has traditionally been the focus of transport studies and urban planning (e.g. Pappalardo et al., 2015; Song et al., 2010).

⁴ There has been some indication that intentions (also expressed in Google searches) are in fact a relevant predictor of future migration both at the national and international levels (e.g. Van Dalen & Henkens, 2013; Tjaden et al., 2019; Böhme et al., 2018).

programming interfaces (APIs). “Big data” includes social media data but is not limited to it. Google, for example, offers a search service which does not operate as a social platform.

Digital trace data are the “results of social interaction via digital tools and spaces as well as digital records of other culturally relevant materials, such as archived newspapers and Google searches including data from popular social networking sites (such as Facebook or Twitter), personal blogs, collaborative online spaces (such as Wikipedia), and data derived from mobile phone or credit card usage” (Cesare et al., 2018: 1980).

The terms “big data” and “digital trace data” refer largely to the same type of sources. However, while term “big data” highlights the *type of* data that is produced, the term “digital trace” data focuses on *how* the data is produced, i.e. through using digital devices (Cesare et al., 2018; Hughes et al., 2016; Sirbu et al., 2020).

New data sources are often collected by private companies for the purpose of offering services to customers. In contrast, “traditional” sources of migration data such as censuses, administrative data and surveys are traditionally collected or made available by government agencies or (publicly funded) research institutes.⁵ This has far-reaching ethical and empirical implications which will be discussed in later sections.

Review criteria

The review will discuss new data sources along five domains: (1) Reliability—the consistency and reproducibility of migration measurements, (2) validity—the accuracy of migration measures and the extent to which data allows to capture the intended concepts used by migration researchers, (3) scope—the breadth and depth of migration-related research that could be explored based on the respective data source, (4) accessibility—the degree to which data is accessible to researchers, and lastly, (5) ethics—the potential risk of violations of data privacy, consent and data protection principles in the data generation process and potential risk of (unintended) harm for research subjects as a result of analysis produced based on new data sources (e.g. Beduschi, 2020; Brayne, 2018; Cesare et al., 2018; Hayes, 2017; Latonero & Kift, 2018; Leese et al., 2021; Molnar, 2019; Zwitter, 2014).

Review of innovative data sources to measure migration

Mobile phones: call detail records and GPS data from smartphone operating systems

Mobile phone Call Detail Records (CDR) can track the approximate location of individuals and, as a result, display movements across space by capturing the call signal sent to cell towers for each outgoing and incoming call (Williams et al., 2015). All caller details are anonymized. Some telecommunications providers are amenable to social research as well, and often provide documented and anonymized digital trace data from their

⁵ Censuses are population counts and are used to calculate ‘stocks’ of migrants (Bilsborrow et al., 1997; Global Migration Group, 2017; White, 2016). Administrative data are collected by governments for the purpose of providing certain services or enforcing certain laws and, thus, may include records of migrants subject to a particular law or service. They may include population registers; visas; immigration, expatriation and asylum records; border entries; work permit registers; tax records, and health insurance or social insurance records (Bilsborrow, 2016; Global Migration Group, 2017; Lemaitre, 2005; Poulain et al., 2006; Willekens et al., 2017). Sample surveys are a common approach to measuring migration, especially in research, given its flexible choice of location, target group and thematic scope (Bilsborrow, 2016; Bilsborrow et al., 1997; Fawcett & Arnold, 1987; Goldstein & Goldstein, 1981).

customers to researchers interested in analysing these data (e.g. Cesare et al., 2018; Chi et al., 2020).

Reliability CDR provides reliable measures of migration in terms of consistency over time. Movements are recorded automatically as required by operating the telephone network. An advantage in terms of reliability is that the information on location does not rely on self-reports by individuals, which may be subject to response biases (a common issue in surveys). However, reliability issues may apply when using CDR data from different operators. This is a common issue because most countries have several telecommunication companies. As consumers switch services, measures of movement over time become less reliable. As a result, CDR is often used on narrowly defined locations and limited time frames.

Validity The key disadvantage is that such data refers to mobile devices, not individuals as such. It is possible that individuals will share the same device, or gift it to others. Furthermore, many migrants may change devices and/or SIM cards after migrating to other countries, given that service providers offer deals limited to particular countries. Therefore, most contributions using phone data have analysed mobility within narrow geographic units (cities, regions) rather than movements across borders. Furthermore, CDRs can be biased because locations are only recorded when calls are made leaving blank spots in the migration process.⁶

Scope While CDR data are usually more helpful for identifying internal (sub-national) migration patterns,⁷ in some cases they can also be used to measure international migration at the sub-regional level, particularly when combined with other sources. For example, CDR have been used to track internal displacement following natural disasters such as the Haiti and Nepal earthquakes (Bengtsson et al., 2011), and the combination of CDR with satellite data can help to map movements between cross-border communities (Hughes et al., 2016). Recent work has leveraged Google Location History data for analysis on migration flows. Google Location History is collected through smartphones that operate the Google Android system and through Google services used through smartphones (e.g. Google Maps or Gmail). Pilot research suggests that this novel source of information could provide information about international migration through ‘fine scale mobility with rare, long distance and international trips’ documented through changes in location by users (Ruktanonchai et al., 2018). Using the same data, Kraemer et al. (2020) described ‘global human mobility patterns, aggregated from over 300 million smartphone users’. According to the authors, the data cover nearly all countries and 65% of earth’s populated surface, including cross-border movements and international migration. The advantage of CDR and location data through smartphone use for measuring migration is the timeliness and detail regarding the location. As such, phone records are particularly useful for studying sudden movements in defined geographic locations. Fast evolving migration situations are difficult to capture with “traditional” data sources such as sample surveys and administrative data, and impossible using censuses.

⁶ Extended Detail Records (XDRs) and Control Plane Records (CPRs) also record locations when something is downloaded or phones are switching antennas providing more complete coverage.

⁷ Measuring human mobility (within countries, regions or cities) is an exploding field of research spanning physics and network science, to data mining, and has fueled advances from public health to transportation engineering, urban planning, official statistics and the design of smart cities (see e.g. Song et al., 2010; Pappalardo et al., 2015).

Without linking mobile phone records to other data sources, CDR provides a limited scope for migration scholars. The only information available is time and location. The type, channel and motivation for a change in location remains unobserved. It thus remains unclear who moved, why people moved, where they wanted to go, who they travelled with, through which channel they travelled, and whether they are likely to stay in their current location. This lack of context information is a key shortcoming compared to “traditional” sample survey research.

Accessibility CDR data is not commonly available to researchers and access depends on willingness of telecommunication companies to collaborate. Different operators in different countries may need to comply with different data protection legislation limiting the extent and level of detail of data that can be shared. In addition, access is often tied to large fees.

Ethics CDR data poses serious ethical concerns. When entering a mobile phone contract or installing a smartphone operating system, many users may not be aware that their location data is collected and analysed for various purposes (Beduschi, 2020; Brayne, 2018; Molnar, 2019). Such data uses are often hidden in the fine print. Since telecommunication and smartphone operating system providers are often private companies, there is a lack of transparency of what companies do with the data. In many countries, governments can mandate companies to provide access to data, for example, for the purpose of criminal investigations (Brayne, 2018). In the field of migration, the granular CDR data can be used to target humanitarian assistance to specific populations in specific locations, however, it could also be used by authorities for enforcing immigration policies, border protection and identifying individuals entering or residing in a country with an irregular status.

Social media

Geo-located social media activity, such as activity on Facebook (Zagheni et al., 2017), Twitter (Chi et al., 2020; Fiori et al., 2017; Martin et al., 2020; UN Global Pulse, 2017; Zagheni et al., 2014), Skype (Kikas et al., 2015), or LinkedIn (State et al., 2014), have been used to infer migration flows and stocks based on the location where users log in or information on location provided by the users themselves through geo-tagged posts or profile information (e.g. nationality or birthplace).

Movement is usually inferred based on changes users make to their self-reported location on the respective platform, or changes in location of log-ins. For example, data from the Facebook advertising platform can yield information on ‘home country’ and country of current residence. This means that Facebook could be used as a ‘real-time census’ to estimate, among other things, the number of users classified by the social media platform as ‘expats’ (users living in a country other than their ‘home country’) at the national or global level at a certain point in time (Zagheni et al., 2017). Using changes in Facebook users’ locations over time, others have identified the increase in the number of Venezuelan migrants in Spain in early 2018, confirmed by official statistics from the Spanish National Statistical Office (Spyratos et al., 2019).

Reliability There are a host of reliability concerns involved in measuring migration using social media data. First, certain segments of the population may be over- or under-represented (for instance, on average, young people are more likely to use Facebook than

older people).⁸ Second, even frequent users may choose not to provide information on their past and current location. Certain types of migrants may deliberately avoid providing information on their location on social networks. Third, it is difficult to verify whether changes in location are accurate, given that this information is sometimes self-reported on a voluntary basis. Fourth, the user base of social media providers constantly changes, which complicates analysis of trends over time (see e.g. Cesare et al., 2018).

Validity With many kinds of social media data, there is a lack of transparency on key measures relevant for migration are generated. For example, there is limited information on how Facebook identifies who is an “expat” or how it labels users as speakers of a different language. This complicates meaningful interpretation of migration patterns observable in the data.

Scope The advantage of geo-located social media data is that, in many countries, certain social media platforms are wildly popular, so that real-time data on large volumes of movements can potentially be accessed. Such data may be particularly useful to study broader migration trends. The level of detail provided by geo-coded social media data is limited in many cases but more extensive compared to CDR data. For example, Facebook provides aggregate-level information on the number of users with specific characteristics such as age, gender, or even education and income proxies as well as a vast range of preferences (measured via users’ “likes” of particular pages). Changes in the characteristics of the number of people living in a specific place are used by researchers to infer ‘migration flows’, assuming that changes in the ‘stock’ of people that report that they live somewhere necessitates that people moves from countries with lower stock numbers to countries with higher stock numbers. Information on friendship networks across countries—recently made available by Facebook—may be used in the future to forecast cross-country migration trends (Tjaden et al., 2021). Despite availability of additional characteristics, the data provide no information about the causes, means, or consequences of migration. There are attempts by governments, law enforcement agencies, international organization and research institutes to monitor the social media activity of migrants before, during and after migration to understand changes in migration patterns (Brenner & Frouws, 2019; Dekker et al., 2018; Sanchez et al., 2018).⁹

Access Many social media companies offer public APIs to allow access to certain parts of their data to researchers. In many cases (e.g. Facebook, Twitter), access can be obtained at no cost which is a substantial advantage over traditional sources such as censuses, administrative data and survey. However, access modalities can change at any given time because data is provided by private companies, rather than taxpayer-funded government or research bodies that are mandated to provide systematic data over time.

Ethics Users of social media are often unaware of the data that is being collected about them and there is a general lack of understanding how such data is and can be used

⁸ Researchers are working to address the methodological challenges associated with such (‘self-selection’) bias, and results look promising (Spyratos et al., 2019; Zagheni et al., 2017; Hughes et al., 2016).

⁹ Europol, Frontex and EASO as well as EU member states have been actively been monitoring communication in Facebook groups relevant for organizing migration, for example, from the Middle East and Africa to the EU. Information on prices of forged documents or sea transport offered by smugglers provide indication on changes in the volume and direction of migration flows. EASO was stopped collecting the data by the EU’s own data protection watchdog, see <https://euobserver.com/investigations/146856>. In theory, it is possible to collect information on users’ profiles and monitor where those users report their location online a year later. This controversial approach has been tried successfully, yet no public evidence has been released.

by companies themselves or third parties (Cesare et al., 2018; Zwitter, 2014). Migration enforcement agencies may use such data for surveillance purposes, which are particularly serious in contexts of irregular migration and forced displacement. Agencies could monitor communication of specific groups or individuals on Twitter and Facebook to identify irregular migrants and track them during their journeys. Companies such as Facebook, however, only allow access to anonymized, aggregate level data to researchers which limits the possibility of using data to harm individuals. Any information on narrowly defined locations and groups becomes inaccessible if the underlying target population decreases beyond a threshold that risks identifying any specific individuals. However, this does not apply to attempts to monitor public communication in social media groups indicating changes in migration patterns. The European Asylum Support Office (EASO) has suspended its efforts to monitor communication of migrants on social media following concerns by the EU own data protection body.¹⁰

Email IP addresses

Repeated logins to the same website and IP addresses from e-mail activity have also been used to estimate international mobility patterns and users' likelihood to move to another country (Zagheni & Weber, 2012). Rather than self-reported location by the user, certain online services such as email providers collect data on where users log into their accounts.

Reliability and validity The same limitation in terms of reliability and validity apply compared to social media data. Similar to log-ins to social media, log-ins to emails are usually recorded via devices (IP addresses) not necessarily people. For example, it is possible—yet presumably rare—that various people use the same email account which will distort any aggregate measure of migration.

Scope The scope of potential migration analysis is further reduced in the case of email log-ins given that additional socio-demographic and socio-economic information about the users (which are available for Facebook) is lacking or not publicly accessible.

Access Most email providers also do not provide public APIs that make data available to researchers. Email communication is considered personal and private communication whereas some communication on social media platforms is (intentionally or unintentionally) made public by users.

Ethics Similar to social media data, there are issues concerning consent and data privacy. Users may not be aware that email providers track their location. In other cases, governmental enforcement agencies may mandate companies to share content of emails of specific individuals for the purpose of criminal investigations or intelligence which bears the potential for misuse also in case of migrants in irregular settings (Brayne, 2018).

Online search data

Online search data has also been used more recently to study migration. Records on Google searchers, for example, have been explored to forecast the number of arrivals of

¹⁰ <https://euobserver.com/investigations/146856>

asylum-seekers in Europe (Connor 2017) or internal migration within the U.S. (Lin et al., 2019). Search data generated through Google's online search platform for migration can be exploited to measure migration intentions and predict subsequent emigration flows (Böhme et al., 2018; UN Global Pulse, 2014). For example, researchers retrieve data on how many times individuals in country A have 'googled' a term that the researcher believes to indicate an intention to migrate (to country B)—for example, 'jobs', 'visa', or the name of the destination country.

Reliability Google Searches are recorded consistently and provide high reliability in terms of the measure as such. The main advantage to Search Data is that Google's search engine is widely used across the globe and has been successfully used to study other social behaviours (e.g. flu outbreaks). Despite broad coverage, important countries (i.e. China) are missing entirely. Issues of reliability emerge regarding applicability across various country contexts, languages and specific populations. Preliminary research in this area suggests that online searches (e.g. via Google searchers) are related to actual movement at the aggregate level, yet the selection of specific search terms in various country contexts appears to be highly important. Syrians looking for ways to flee to Europe 'google' different terms than Canadians looking for a job in the US. The meaning of the same search terms may also vary in different languages. Overall, this means that Google searches may be indicative of migration from a certain country to another country, but difficult to scale up to multiple migration contexts (see Tjaden et al., 2021).

Validity Online search data has one obvious shortcoming: 'searching' is not 'doing'. Just because someone looks up information on another country or, more explicitly, gathers information on how to move to another country, does not mean that they will actually move. Search data (similarly to survey data on emigration intentions) are a 'pre-behavioural' proxy for actual migration. Some studies suggest that intentions are a good predictor for eventual migration (Van Dalen & Henkens, 2013; Tjaden et al., 2019), but research also suggests that the strength of the predictor varies considerably based on where migrants are from and where they want to go (Tjaden et al., 2019).

Scope A major disadvantage of Google search data is the high level of aggregation at which data is made available. Search data is made available at the population level for countries or, in certain countries like the US, for subregions. Search data does not include any additional information about those who show interest in migrating, and thus renders any individual-level analysis impossible.

Access Google search data is freely and publicly accessible via the Google Trends platform and API.

Ethics The potential risk of misuse of data is limited given the high level of aggregation and anonymity of data which the company makes available. Serious concerns would arise when data for specific locations and IP addresses is used to infer individual level migration behaviour. Google itself is analysing individual-level location data to provide targeted advertisements to users who use their search engine. However, there is a lack of transparency in terms of the conditions under which such data may be shared with governments or other third parties. In addition, usual concerns around unawareness among users about the usage of their data apply.

Bibliometric data

Bibliometrics is a field of research that uses statistical methods to systematically analyse publications records (books, articles etc.). One sub-field of bibliometrics—scientometrics—is the analysis of scientific publications. Detailed information about academic output is recorded and made accessible through scientific databases (e.g. Scopus, Web of Science, Google scholar and others). This information has been used to model the international mobility of academics (Czaika & Orazbayev, 2018; Laudel, 2003; Moed & Halevi, 2014; Sudakova & Tarasyev, 2019; Wang et al., 2019). Changes in the researchers' affiliation to institutions located in different countries indicates migration.

Reliability Measuring migration through changes in affiliations is consistent and reliable. Scientists have an interest to publish their work in recognized journals and books, institutions have an interest that researchers indicate their home institution, and most research outlets make it mandatory for authors to provide this information. Nevertheless, the data is sensitive to the accuracy of self-reported data which can be outdated.

Validity Migration analysis based on bibliometric data has the potential to collect additional context information including socio-demographic characteristics of the professionals (age, gender, ethnic origin, for example, may be inferred based on name recognition algorithms and web scraping individual professionals' web pages). Additional information about the universities, faculty and chair may be matched with additional effort.

Scope The drawback of this data source is its restriction to a narrowly defined group of professionals (i.e. academics) where public access to their affiliation is the norm. However, it may be possible to extend this approach to other fields of professionals where public information on affiliations is common (i.e. athletes, musicians etc.).

Access Bibliographic data has become available through the digitalization of entire libraries, records of publishers, academic journals, and ambitious projects such as Google Books and Google Scholar that aim to record any academic publications that is published. Most academics provide their affiliations publicly to gain visibility and broaden their reach.

Ethics Compared to previously described sources, ethical concerns are limited because the personal information used for analysis is provided voluntarily and knowingly. The population is restricted to regular labour migrants which limits the potential for misuse by authorities.

Remote sensing technologies

Remote sensing is an umbrella term for collecting information about something without making physical contact. In current usage, remote sensing refers to the use of satellite or aircraft-based sensor technologies (i.e. drones). Remote sensing is commonly used in geography, earth sciences, climate research, agricultural studies, wildlife studies, military, and intelligence gathering, but also increasingly for urban planning, tourism, commerce, and various humanitarian applications (Miller et al., 2019). Changes in human activity visible in the images (i.e. settlements, refugee camps, light emissions at night) can be used to infer mobility.

Reliability If applied consistently, the approach to measuring migration using remote sensing technology by averaging physical quantities over pixels can yield reliable migration measures. Algorithms automatically detect changes in visual patterns on satellite or drone images over time. For example, the population size of settlements can be estimated by counting rooftops visible on satellite/drone images. Depending on the proximity and resolution of the image, individuals within certain localities can be identified. Comparing images over time can be used to estimate immigration and emigration into a certain, narrowly defined, location.

Validity The obvious downside of satellite and drone images for measuring migration is that no additional individual-level information about migrants is available: Who is moving, from where, to where, how etc. By itself, remote sensing provides information on how many tents, rooftops or individuals are present in a certain locality, but no information about what happened when there are less dots and shadows the next time new images become available.

Scope There is a rapidly growing number of examples with relevance for migration studies. First, drones and satellite images inform policies and direct aid to refugees. For instance, the United Nations Institute for Training and Research (UNITAR) mapped refugee camps in Jordan and elsewhere with its Operational Satellite Applications Programme.¹¹ Civil society organizations such as Human Rights Watch or Amnesty International use satellite imagery to document humanitarian needs of displaced populations at borders or in refugee camps by measuring the growth of settlements.¹² In this case, satellite images are providing an indication of where aid and assistance are most needed (Bitelli et al., 2017; Quinn et al., 2018; Shatnawi et al., 2020; Tiede et al., 2017).

Satellite imagery also forms a key part of the ‘smart border’ agenda, which attempt to use modern technology to improve border management around the world and track ‘illegal’ crossings. Systems relying on remote sensing were developed “to assist border authorities with more effective surveillance and reliable decision-making support” (Al Fayez et al., 2019). In contrast, civil society organizations use the same technology to monitor deaths and violations of migrants’ rights at the maritime borders of the EU.¹³

For the moment, remote sensing appears to be most useful for informing operations on the ground (managing refugee camps, targeting humanitarian assistance, managing borders etc.) and less for research on migration per se. The technology can also be used to monitor slow onset emigration rates due to changes in climate which can also be inferred from images.

Access With improvements in the quality and accessibility of satellite imagery (Popkin, 2018) provided by the European Space Agency, NASA, and others, researchers are also

¹¹ <https://unitar.org/sustainable-development-goals/satellite-analysis-and-applied-research>

¹² <https://earthispace.blog/satellite-solution-refugee-crisis/>
<https://www.hrw.org/news/2016/09/07/jordan-new-satellite-images-syrians-stranded-border>
<http://www.satsentinel.org/>
<https://www.unglobalpulse.org/2020/02/pulsesatellite-a-collaboration-tool-using-human-ai-interaction-to-analyse-satellite-imagery/>

¹³ For example, governments use drone imagery to monitor immigration attempts in the Mediterranean Sea and at Europe’s land borders (Bhadwal et al., 2019; Dijkstra et al., 2017) or to monitor border crossing attempts at the Mexican-US border (see <https://www.gao.gov/assets/690/682842.pdf>). Greek and EU authorities use satellite imagery to monitor crossing attempts (ibid.). NGOs and universities also use the same information to document human rights abuses by, for example, using satellite imagery and other evidence to reconstruct the journey of a boat with migrants that lost dozens passengers upon its return to Tripoli (Dijkstra et al., 2017). Although the boat had been spotted by several aircraft and vessels, no rescue operation had been mounted.

exploring ways to use remote sensing data to measure human migration globally. Public and private bodies offer access to satellite imagery for research purposes and tech companies offer cloud computing power to conduct complex and demanding analyses within minutes.¹⁴ Depending on the specific data provider, access can be free of charge to research institutes or come with a fee.

Ethics Ethical issues are a key concern for remote sensing technologies because information is collected without the knowledge or consent of individuals. New high-resolution satellite imagery and drone images can identify individuals using face recognition technology. Law enforcement, policing and intelligence agencies use such approaches (Brayne, 2018; Hayes, 2017; Leese et al., 2021; Molnar, 2019) which raises serious concerns regarding the situation in undemocratic countries with low data protection standards and policies aiming to suppress and control groups in society. Drones may also be increasingly operated by companies in addition to governments which raises concerns over unknown privacy violations by non-governmental actors.

International air travel

Upon first view, international air passenger traffic belongs to the realm of tourism and transport studies, not migration (see Sect. 2.1.). However, there have been attempts to use this information to infer migration flows. For example, Gabrielli et al. (2019) used dyadic monthly air passenger traffic between 239 countries and territories worldwide from January 2010 to March 2018 to estimate the number of passengers on commercial flights operated globally. The study explored whether a surplus in travel (increase in travel from A to B but no increase in return travel from B to A within a year) can be linked to migration flows.

Reliability Air passenger data is highly standardized and consistent as it is subject to international industry standards.

Validity Passenger data does not measure migration directly and can only be used to infer different types of migration by inference. Since air passengers data does not allow to track individual passengers or specific cohorts on the basis of their date of entry, researchers need to make assumptions about the length of stay of the passengers. This is problematic because the publicly available data does not indicate who the passengers are, how long they will stay in the country, on which visa they are travelling etc. In addition, flight passenger data is a selective picture of global mobility. 44 percent of registered cross-border travels occur through commercial flights, and that this proportion increases at rising distances between countries (Recchi et al., 2019).

Scope Overall, the data may be used to estimate international migration flows if combined with additional data sources. At the moment, the research is still in its exploratory stage and the methods appear underdeveloped. In the future, this approach may bear the potential to measure the level of visa overstayers between countries, one indicator of irregular migration.

¹⁴ See for example <https://earth.esa.int/eogateway/>; <https://search.earthdata.nasa.gov/search>; <https://earthexplorer.usgs.gov/>

Access Air passenger data is collected by flight companies which some make available for purchase. The EU recently made a public and free dataset available.¹⁵

Ethics Ethical concerns regarding flight data are limited in its current state of the available data. Flight data is aggregated at the country and month level and anonymized. Currently, any misuse for the disadvantage for individuals is unlikely.

Online news data

New advances in technology have made available online news aggregators such as Google News or the Global Database of Events, Language, and Tone (GDELT).¹⁶ Such platforms monitor the world's news media from nearly every corner of every country in print, broadcast, and web formats. This data has the potential to capture acts of past or prospective migration that were not covered in traditional sources such as administrative data or surveys.

Reliability Migration measures based on online news aggregator data can be considered reliable to the extent that algorithms deriving information on migration apply consistently across all countries and news sources. The issue is that the success of the algorithm in detecting migration may vary by country, by quality of the news outlets, by language and type of migration to be covered. In addition, algorithms may capture the same migration events several times as the same event may have been covered by several news outlets.

Validity The large volume of news articles required to collect information on migration encourages researchers to use language processing algorithms. The emerging evidence is still unclear on how accurately such algorithms may detect events that actually capture migration.

Scope Approaches are still very recent but several uses of this data are available. Carammia et al., (2020) have used the GDELT database to measure political, social, economic “push factor” events that could motivate people to leave their country. In combination with other data sources, they attempt to forecast displacement and migration with a view to set up early warning systems currently under development in the EU.¹⁷ In a similar vein, the Internal Displacement Monitoring Centre (IDMC) uses the GDELT database to track internal displacement.¹⁸ The IOM is also experimenting with such data to improve analysis of the number of migrants that went missing along their journeys, such as the IOM's missing migrants project (Borja & Black, 2021). Apart from eye-witness reports, news articles are the main way to systematically collect data on migrant fatalities and bring light to this tragic topic for policymakers.

Access GDELT and Google News can be accessed online free of charge for researchers.

Ethics Ethical concerns arise in countries with low standards of journalism and data privacy. It is possible, for example, that the identity of individual migrants is revealed

¹⁵ See the EU Knowledge Centre for Migration and Demography's Dynamic Data Hub, at: <https://bluehub.jrc.ec.europa.eu/migration/app/?state=5d6005b30045242cabd750a2>.

¹⁶ GDELT is a repository of 316 types of geolocated event reported in the world's broadcast, print and web media, in 100 languages.

¹⁷ Frontex, EASO and JRC are all developing early warning, forecasting and foresight approaches.

¹⁸ <https://www.internal-displacement.org/expert-opinion/identifying-how-technology-and-collaboration-between-innovators-can-help-ensure-no-one>

in a news article and picked up by automated text analysis. In theory, this information could be used by enforcement agencies to press charges in case of irregular migration or used by smugglers for debt collection. Even when interviewees provide consent for their personal information to be used, they may not be aware that their information may enter migration databases. Such abuses are possible with traditional media sources, however, digital applications may exacerbate the problem by providing cheaper, faster and broader access to data.

Discussion and conclusion

This review highlighted both the enormous opportunity of “big data” and “digital trace data” to complement traditional sources of migration data (see Cesare et al., 2018; Hilbert, 2016; Hughes et al., 2016; Laczko & Rango, 2014; Rango & Vespe, 2017; Sirbu et al., 2021) and the main challenges and risks associated with such data. Several broader conclusions can be drawn from the above discussion on the eight discussed sources (see Table 1 for a summary):

The main advantages of digital data sources for migration scholars are captured by the first two V's of the “3 V's definition” of big data introduced in Sect. 3.2: Volume and velocity. In some cases, digital data sources provide information of millions of individuals in almost real-time. This provides migration researchers with the possibility explore migration trends where administrative data sources, surveys and censuses (sources traditionally used for inferring migration) are not available (such as in many low-income contexts), not accessible or too slow (for example in contexts of displacement and forced migration that are unfolding rapidly). Another key advantage is its granularity by providing “high-resolution” information. Digital data sources, especially approaches leveraging remote sensing technologies or mobile cell phones, often allow researchers to zoom into migration events at the sub-national, sub-regional or even local level. Nationally representative survey data, for example, often lacks sufficient sample size to disaggregate to the level of regions, districts or cities.

Unlike “traditional” sources, new data sources often make no distinction in terms of the “legal residence status” of individuals. Anyone using a digital service is included in the data. As a result, the volume, granularity and status-agnostics of many digital data sources offer new opportunities to collect information on “hard-to-reach” populations such as recent migrants, displaced or forced migrants and irregular migrants which are often excluded from ‘official’ data sources such as population registers or surveys (Cesare et al., 2018; Massey & Capoferro 2004; Reichel & Morales, 2017). Lastly, a key advantage to migration scholars is the fact that many data sources are accessible online and free of charge.

The review has also highlighted major challenges associated with using digital data sources for inferring migration. ‘Digital trace’ data is largely collected by private companies who offer services to their users and use user data to target advertisements or sell data to third parties. These data are not designed for research purposes. This has important implications both on ethical (Beduschi, 2020; Brayne, 2018; Cesare et al., 2018; Hayes, 2017; Latonero & Kift, 2018; Leese et al., 2021; Molnar, 2019; Sirbu et al., 2020; Zwitter, 2014) and empirical grounds in terms of reliability and validity of migration measurement (Cesare et al., 2018; Ruel et al., 2016; Sirbu et al., 2020).

Table 1 Overview of measurement approaches

#	Data source	Provider	Main approach to inferring migration stocks or flows	Example of applications to migration studies
1	Mobile phones [CDR, GPS; Android location history]	Private sector companies	Tracks location of individuals via phone signals through cell phone towers or by storing GPS location through the smartphone operating system	Displacement following disasters; internal mobility at local level
2	Social media [e.g. Facebook, Twitter, Skype, LinkedIn]	Private sector companies	Measure changes in users' self-reported location in their profiles or posts and the location of IP addresses from where users log into their accounts	Estimating international migration stocks and flows; Estimating changes in friendship networks, monitoring migration-related communication in social media groups
3	Online search data [Google, Bing]	Private sector companies	Measuring popularity of migration-relevant search terms online	Forecasting potential bilateral migration flows
4	IP addresses [email logins]	Private sector companies	Track locations from which users access their email account	Estimating international migration flows
5	Bibliographic data	Companies; Academic bodies	Tracking changes in the publicly available institutional affiliation of professionals (i.e. academics)	Measuring international migration of academics
6	Remote sensing [satellite and drone images]	Government; private sector companies	Measuring changes in physical features (rooftops, settlements, light emission at night) in defined locations on maps recorded via satellite or drone imagery	Measuring rapid changes in sizes of settlements and refugee camps; border management; targeting humanitarian assistance
7	Air passenger data	Private sector companies	Measuring discrepancies in the volume between flight arrivals and returns as a proxy for immigration	Estimating international migration flows
8	Online news platforms	Private sector companies; research institutes	Compiling information on mentions of (past or potential) migration events in news media	Estimating internal displacement; estimating forced migration, estimating migrant fatalities

First, there are severe ethical concerns regarding the use of digital data sources including the limited awareness of users regarding the extent of and purposes for which their data is being used and the risk of harm for individual migrants in cases where information is used by law enforcement, border management, intelligence agencies or smugglers (Brayne, 2018; Hayes, 2017; Molnar, 2019). The first rule that a researcher must follow is to acknowledge that data are people and can do harm (Sirbu et al., 2021). Data may include information on particular vulnerable groups such as refugees, irregular migrants, persons displaced by disasters. Some may be persecuted by authorities in origin and destination countries. Researchers must ensure ethical standards for data use that protect vulnerable groups from identification and possible discrimination (Cesare et al., 2018).

Violations of data privacy and protection standards are especially concerning in undemocratic countries with low data protection standards, limited rule-of-law, and a lack of democratic norms. In extreme cases, new technologies can enable “digital authoritarianism” and “Orwellian state surveillance” (Dragu & Lupu, 2020). Three examples illustrate the extent of real risks: China’s social credit system, for example, leverages smartphone location data, social media communication, travel records, purchase records, camera data and facial recognition, among others, in combination with various administrative records to assign a social credit to its citizen. Low scores could be used to prevent access to a passport or visa needed to leave the country.¹⁹ After departure of US troops, the Taliban have reportedly considered using US-made digital identity technology to persecute Afghans who have worked with the international coalition. Funded by millions of donor funding, Afghanistan’s National Statistics and Information Authority launched a digital biometric identity card including fingerprints, iris scans and a photograph, as well as voter registration databases.²⁰ In 2018, Bangladesh shared hundreds of thousands of data of Rohingya refugees collected by UNHCR with Myanmar then used to facilitate potential repatriation.²¹

Unlike states, researchers often do not have access to individual-level information from digital sources, however, they must be aware of the potential harms for individuals and groups when using digital data sources and must review the data providers’ data protection and privacy standards. Researchers are advised to seek ethical approval by a scientific committee when dealing with digital data and migration.

Apart from ethics, the review has highlighted many empirical challenges. Reliability and validity of digital data sources for inferring migration must be considered when engaging in research. It is often not transparent how exactly key measures of migration are generated (i.e. Facebook, Google). There are also concerns regarding “generalizability” of digital data as the user base of various digital services is often selective and

¹⁹ The system has reportedly been used to suppress and control the Uighur minority in China’s Xinjiang region. Analysis to linked data from various digital data sources alerts enforcement agencies at checkpoints when Uighurs approach the limits of their neighbourhood. (see Ross Andersen in the Atlantic: “The panopticon is already here”, available at <https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/>)

²⁰ See Emrys Schoemaker, 7 September 2021 in the Guardian: “The Taliban are showing us the dangers of personal data falling into the wrong hands”, available at <https://www.theguardian.com/global-development/2021/sep/07/the-taliban-are-showing-us-the-dangers-of-personal-data-falling-into-the-wrong-hands>

²¹ See Human Rights Watch on 21 June 2021: “UN Shared Rohingya Data Without Informed Consent”, available from <https://www.hrw.org/news/2021/06/15/un-shared-rohingya-data-without-informed-consent>

does not represent the general population at large (Cesare et al., 2018; Sirbu et al., 2020). Digital data is often made available at highly aggregated levels. This severely limits the analytical potential for migration scholars who are often interested in measuring migration at the individual (micro) level. Moreover, digital data often remains “thin” offering very few additional information beyond time and location such as socio-demographic or socio-economic characteristics of migration, the reasons and channels of migration or the expected duration of stay. This further limits its analytical use especially if the data is not combined with other data sources. Traditional surveys usually do not face these issues (but many others) as they are tailored for answering specific research questions.

Lastly, many populations are excluded from a variety of digital data sources despite technological advances worldwide. For those data relying in digital traces such as social media or online searches, it is still a long way to obtain a comprehensive picture of global mobility as smartphone user penetration reached 38.5 percent in 2020 and half the world population is still offline today.²²

Looking ahead, several trends are already unfolding. First, new breakthroughs in measuring migration research will stem from a combination of different sources (e.g. Alexander et al., 2020; Sirbu et al., 2020; Snijders et al., 2012). Given the ‘representativeness’ issue of digital trace data, traditional data is needed for “ground truthing” (i.e. cross-validating data by comparing it with other ‘official’ data sources). To address, the “thinness” of digital data, combining data provides large opportunities to add richness to the analysis. One example is using social media, online search, or mobile phone data to locate migration events or patterns and then target surveys in certain geographies or adjust administrative data collection accordingly (Alexander et al., 2020).

The second trend is further convergence and integration of academic disciplines around the issue of measuring migration (e.g. Miller et al., 2019). Different disciplines such as earth sciences, climate research, security studies, tourism and transport studies, computer science, sociology, economics, demography, ethnography, library sciences and political science bring different tools, methodologies and technologies to the table which will likely see the field become even more interdisciplinary. As a result, more interdisciplinary dialogue is needed to advance the field.

As the field is changing at a dizzying speed, this paper attempted to provide a brief overview and reflection on the main new digital data sources. The aim of this review was to provide incoming migration researchers with a menu of options and seasoned researchers with an update on new approaches. The information provided assists researchers in making difficult trade-offs when approaching their research question and policy analysts with a broad understanding of the limitations of the data they use.

The review has two obvious limitations. First, given the complex and rapidly growing field it seems near impossible to cover every existing approach and to cover the literature in all its breath. The review is focused on the main approaches without any claim to be exhaustive. Second, the field is rapidly evolving. This means that new research will have become available already by the time of publication.

²² <https://data.worldbank.org/indicator/IT.NET.USER.ZS>

The toolbox for migration researchers will become bigger, more diverse, but also more powerful due to new opportunities of digital data. Despite the ‘gold rush’ on big and digital data, the review also cautioned migration scholars in view of the many ethical and empirical obstacles for inferring migration based on digital data sources. The review aimed to contribute to a balanced understanding of these new data sources to facilitate knowledge accumulation and interdisciplinary dialogue in the field of migration studies.

Acknowledgements

The author would like to thank Marzia Rango (IOM), Emilio Zagheni (Max Planck Institute for Demographic Research), Ingmar Weber (Qatar Computing Research Institute) and Teddy Wilkin (EASO) for numerous discussions and presentations on the potential of digital trace data for measuring migration which motivated the idea for this paper.

Authors' contributions

Single authorship/ not applicable. Author read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

Not applicable.

Declarations

Competing interests

The author confirms no competing interests.

Author details

¹Department of Economics and Social Sciences, University of Potsdam, August-Bebel-Str.89, 14482 Potsdam, Germany.

²International Organization for Migration's (IOM) Global Migration Data Analysis Centre (GMDAC), Berlin, Germany.

Received: 21 October 2020 Accepted: 2 November 2021

Published: 21 December 2021

References

- Al Fayed, F., Hammoudeh, M., Adebisi, B., & Abdul Sattar, K. N. (2019). Assessing the effectiveness of flying ad hoc networks for international border surveillance. *International Journal of Distributed Sensor Networks*, 15(7), 1–12.
- Alexander, M., Polimis, K., & Zagheni, E. (2020). Combining social media and survey data to Nowcast migrant stocks in the United States. arXiv preprint <https://arxiv.org/abs/2003.02895>.
- Beduschi, A. (2020). International migration management in the age of artificial intelligence. *Migration Studies*. <https://doi.org/10.1093/migration/mnaa003>.
- Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS Medicine*, 8(8), e1001083.
- Bhadwal, N., Madaan, V., Agrawal, P., Shukla, A., & Kakran, A. (2019). Smart border surveillance system using wireless sensor network and computer vision. In *2019 international conference on Automation, Computational and Technology Management (ICACTM)* (pp. 183–190). London. <https://doi.org/10.1109/ICACTM.2019.8776749>.
- Bilsborrow, R. E. (2016). Concepts, definitions and data collection approaches. In M. J. White (Ed.), *International handbook of migration and population distribution* (pp. 109–156). Springer.
- Bilsborrow, R. E., Hugo, G., & Oberai, A. S. (1997). *International migration statistics: Guidelines for improving data collection systems*. International Labour Organization.
- Bitelli, G., Eleias, M., Franci, F., & Mandanici, E. (2017). VHR satellite imagery for humanitarian crisis management: A case study. In: *Fifth international conference on remote sensing and geoinformation of the environment (RSCy2017)*, 10444, p. 104440T. International Society for Optics and Photonics.
- Böhme, J., Gröger, A., & Stöhr, T. (2018). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142, 1–14.
- Borja, A. G., & Black, J. (2021). Measuring migrant deaths and disappearances. *Forced Migration Review*, 66, 58–60.
- Brayne, S. (2018). The criminal law and law enforcement implications of big data. *Annual Review of Law and Social Science*, 14, 293–308.
- Brenner, Y., & Frouws, B. (2019). Hype or hope? Evidence on use of smartphones & social media in mixed migration. Mixed Migration Centre, Geneva.
- Carammia, M., Iacus, S. M., & Wilkin, T. (2020). Forecasting asylum-related migration flows with machine learning and data at scale. arXiv preprint [arXiv:2011.04348](https://arxiv.org/abs/2011.04348).
- Cesare, N., Lee, H., McCormick, T., Spiro, E., & Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography*, 55(5), 1979–1999.
- Chi, G., Lin, F., Chi, G., & Blumenstock, J. (2020). A general approach to detecting migration events in digital trace data. *PLoS ONE*, 15(10), e0239408.

- Clemens, M., Summers, L. H., & Santo Tomas, P. A. (2009). Migrants count: Five steps towards better migration data. In *Report of the commission on international migration data for development research and policy*. Center for Global Development.
- Czaika, M., & Orazbayev, S. (2018). The globalisation of scientific mobility, 1970–2014. *Applied Geography*, 96, 1–10.
- Dekker, R., Engbersen, G., Klaver, J., & Vonk, H. (2018). Smart refugees: How Syrian asylum migrants use social media information in migration decision-making. *Social Media+ Society*, 4(1), 2056305118764439.
- Dijstelbloem, H. (2017). Migration tracking is a mess. *Nature*, 543(7643), 32–34.
- Dragu, T., & Lupu, Y. (2021). Digital authoritarianism and the future of human rights. *International Organization*, 75(4), 991–1017.
- Fawcett, J. T., & Arnold, F. (1987). The role of surveys in the study of international migration: An appraisal. *International Migration Review*, 21(4), 1523–1540.
- Fiori, L., Abel, G., Cai, J., Zagheni, E., Weber, I., & Vinué, G. (2017). Using Twitter data to estimate the relationship between short-term mobility and long-term migration. In *Proceedings of the 2017 ACM on web science conference* (pp. 103–110).
- Gabrielli, L., Deutschmann, E., Natale, F., Recchi, E., & Vespe, M. (2019). Dissecting global air traffic data to discern different types and trends of transnational human mobility. *EPJ Data Science*, 8(1), 26.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
- Global Migration Group. (2017). Handbook for improving the production and use of migration data for development. In *Global Knowledge Partnership for Migration and Development (KNOMAD)*. World Bank.
- Goldstein, S., & Goldstein, A. (1981). Surveys of migration in developing countries: A methodological review. *Papers of the East-West Population Institute*, 71, 120.
- 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*, 99(904), 1Y79–209.
- Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, 34(1), 135–174.
- Hughes, C., Zagheni, E., Abel, G. J., Sorichetta, A., Wiśniowski, A., Weber, I., & Tatem, A. J. (2016). *Inferring migrations: Traditional methods and new approaches based on mobile phone, social media, and other big data: Feasibility study on inferring (labour) mobility and migration in the European Union from Big Data and Social Media Data*. European Commission.
- IOM (2020). ‘Chapter 3: Migration Research and Analysis: Growth, Reach and Recent Contributions’. In: *World Migration Report 2020*. Geneva.
- Kikas, R., Dumas, M., & Saabas, A. (2015). Explaining international migration in the skype network: The role of social network features. In: *Proceedings of the 1st ACM Workshop on Social Media World Sensors* (pp. 17–22).
- Kraemer, M. U., Sadilek, A., Zhang, Q., Marchal, N. A., Tuli, G., Cohn, E. L., Hswen, Y., Perkins, T. A., Smith, D. L., Reiner Jr. R. C., & Brownstein, J. S. (2020). Mapping global variation in human mobility. *Nature Human Behaviour*, 4(8), 800–810.
- Laczko, F. (2016). Improving Data on International Migration and Development: Towards a Global Action Plan? Conference Paper, *Improving Data on International Migration – Towards Agenda 2030 and the Global Compact on Migration*, 1–3 December, 2016, German Federal Foreign Office.
- Laczko, F., & Rango, M. (2014). Can big data help us achieve a “Migration Data Revolution”? *Migration Policy Practice (IOM)*, 4(2), 20–29.
- Latonero, M., & Kift, P. (2018). On digital passages and borders: Refugees and the new infrastructure for movement and control. *Social Media+ Society*, 4(1), 1–11.
- Laudel, G. (2003). Studying the brain drain: Can bibliometric methods help? *Scientometrics*, 57(2), 215–237.
- Leese, M., Noori, S., & Stephan S. (2021). Data matters: The politics and practices of digital border and migration management. *Geopolitics*. <https://doi.org/10.1080/14650045.2021.1940538>.
- Lemaître, G. (2005). The comparability of international migration statistics: Problems and prospects. *Statistics Brief*, 9, 1–8.
- Lin, A. Y., Cranshaw, J., & Counts, S. (2019). Forecasting US domestic migration using internet search queries. In: *The world wide web conference* (pp. 1061–1072).
- Martín, Y., Cutter, S. L., Li, Z., Emrich, C. T., & Mitchell, J. T. (2020). Using geotagged tweets to track population movements to and from Puerto Rico after hurricane Maria. *Population and Environment*, 1(1), 1–24.
- Massey, D., & Capoferro, C. (2004). Measuring undocumented migration. *The International Migration Review*, 38(3), 1075–1102.
- Miller, H. J., Dodge, S., Miller, J., & Bohrer, G. (2019). Towards an integrated science of movement: Converging research on animal movement ecology and human mobility science. *International Journal of Geographical Information Science*, 33(5), 855–876.
- Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, 101(3), 1987–2001.
- 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.
- Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., & Barabási, A. L. (2015). Returners and explorers dichotomy in human mobility. *Nature Communications*, 6(1), 1–8.
- Pisarevskaya, A., Levy, N., Scholten, P., & Jansen, J. (2019). Mapping migration studies: An empirical analysis of the coming of age of a research field. *Migration Studies*, 8(3), 455–481.
- Popkin, G. (2018). Technology and satellite companies open up a world of data. *Nature*, 557(7706), 745–748.
- Poulain, M., Perrin, N., & Singleton, A. (2006). *Towards harmonised european statistics on international migration*. Presses Universitaires de Louvain.
- Quinn, J. A., Nyhan, M. M., Navarro, C., Coluccia, D., Bromley, L., & Luengo-Oroz, M. (2018). Humanitarian applications of machine learning with remote-sensing data: Review and case study in refugee settlement mapping. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128), 1–16.

- Rango, M., & Vespe, M. (2017). Big data and alternative data sources on migration: from case-studies to policy support. Summary Report. Ispra: European Commission Joint Research Centre.
- Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W., & Bijak, J. (2013). Integrated modeling of European migration. *Journal of the American Statistical Association*, 108(503), 801–819.
- Recchi, E., Deutschmann, E., & Vespe, M. (2019). Estimating transnational human mobility on a global scale. Robert Schuman Centre for Advanced Studies Research Paper No. RSCAS, 30.
- Reichel, D., & Morales, L. (2017). Surveying immigrants without sampling frames—Evaluating the success of alternative field methods. *Comparative Migration Studies*, 5(1), 1.
- Ruel, E., Wagner III, W., & Gillespie, B. (2016). The quality of measurement: reliability and validity. The practice of survey research: Theory and applications.
- Ruktanonchai, N. W., Ruktanonchai, C. W., Floyd, J. R., & Tatem, A. J. (2018). Using Google Location History data to quantify fine-scale human mobility. *International Journal of Health Geographics*, 17(1), 1–13.
- Sanchez, G., Hoxhaj, R., Nardin, S., Geddes, A., Achilli, L., & Kalantaryan, S. (2018). A study of the communication channels used by migrants and asylum seekers in Italy, with a particular focus on online and social media.
- Shatnawi, N., Weidner, U., & Hinz, S. (2020). Monitoring urban expansion as a result of refugee fluxes in North Jordan using remote sensing techniques. *Journal of Urban Planning and Development*, 146(3), 1–32.
- Sirbu, A., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., & Pappalardo, L. (2020). Human migration: the big data perspective. *International Journal of Data Science and Analytics*, 11(0), 1–20. <https://doi.org/10.1007/s41060-020-00213-5>.
- Snijders, C., Matzat, U., & Reips, U. D. (2012). Big data: Big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, 7(1), 1–5.
- Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in human mobility. *Science*, 327(5968), 1018–1021.
- Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2019). 'Quantifying International Human Mobility Patterns Using Facebook Network Data. *PLoS ONE*, 14(10), 1–22.
- State, B., Rodriguez, M., Helbing, D., & Zagheni, E. (2014). Migration of professionals to the U.S.: Evidence from linkedin data. In L. M. Aiello, & D. McFarland (Eds.), *6th international conference on social informatics*, SocInfo (pp. 531–543). Springer.
- Sudakova, A. E., & Tarasyev, A. A. (2019). Digitalization and scientometrics in assessing the migration of scientists. In *International scientific and practical conference on digital economy* (ISCDE 2019). Atlantis Press.
- Tiede, D., Krafft, P., Füreder, P., & Lang, S. (2017). Stratified template matching to support refugee camp analysis in OBIA workflows. *Remote Sensing*, 9(4), 326.
- Tjaden, J., Arau, A., Nuermairaiti, M., Cetin, I., Acostamadiedo, E., & M. Rango (2021). Using "Big Data" to forecast migration—A tale of high expectations, promising results and a long road ahead'. *MEDIUM*. Available at <https://medium.com/@UNmigration/using-big-data-to-forecast-migration-8c8e64703559>.
- Tjaden, J., Auer, D., & Laczko, F. (2019). Linking migration intentions with flows: Evidence and potential use. *International Migration*, 57, 36–57.
- UN DESA. (2017). 'UN Handbook on Measuring International Migration through Population Censuses'. In: *Economic and Social Affairs*. New York.
- UN DESA. (2019). International Migration 2019. Department of Economic and Social Affairs, Population Division, New York City.
- UN Global Pulse. (2014). Estimating Migration Flows Using Online Search Data. *Global Pulse Project Series*, 4, 1–2.
- UN Global Pulse. (2017) *Social media and forced displacement: Big data analytics & machine-learning*, Geneva.
- UNHCR. (2020). UNHCR Global Trends - Forced displacement in 2020. United Nations High Commissioner for Refugees, Geneva.
- Van Dalen, H. P., & Henkens, K. (2013). Explaining emigration intentions and behaviour in the Netherlands, 2005–2010. *Population Studies*, 67(2), 225–241.
- Wang, Y., Luo, H., & Shi, Y. (2019). Complex network analysis for international talent mobility based on bibliometrics. *International Journal of Innovation Science*, 11(3), 419–435.
- White, M. J. (Ed.). (2016). *International Handbook of Migration and Population Distribution* (vol. 6). Springer.
- Willekens, F., Massey, D., & Raymer, J. (2017). International migration under the microscope. *Science*, 352(6288), 897–899.
- Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., & Dobra, A. (2015). Measures of human mobility using mobile phone records enhanced with GIS Data. *PLoS ONE*, 10(7), 1–16.
- Zagheni, E., Garimella, V. R. K., Ingmar, W., & State, B. (2014). Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 439–44). ACM Press.
- Zagheni, E., & Weber, I. (2012). You are where you e-mail: Using e-mail data to estimate international migration rates. In *Proceedings of the 4th annual ACM web science conference* (pp. 348–351). ACM Press.
- Zagheni, E., Weber, I., & Gummedi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43, 721–734.
- Zwitter, A. (2014). Big data ethics. *Big Data and Society*, 1(2), 2053951714559253.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.