Emerging Big Data Sources for Studying Migration

Tuba Bircan, *Dept. of Sociology, Vrije Universiteit Brussel*

Albert Ali Salah*, Dept. of Information and Computing Sciences, Utrecht University*

Abstract

In this chapter, we provide an overview of migration measures based on traditional data sources and describe the use of emerging big data sources for migration and mobility indicators. We provide applications of big data analytics for migration measures, discussing ethical considerations of using big data in this domain. We do not detail particular artificial intelligence and machine learning approaches used for processing big data, which are necessary to deal with large data volumes, and particularly useful for prediction purposes, but focus rather on the advantages and disadvantages of the specific data sources and provide example applications. At the end of the chapter, we speculate about what we can expect in the coming years, in terms of advancing artificial intelligence approaches and the efforts to regulate this difficult-to-check growth, but also in terms of increased automated surveillance in every sphere of society.

Note: This is the uncorrected author proof. Copyright with Edward Elgar. https://www.e-elgar.com/shop/gbp/handbook-of-research-methods-in-migration-9781800378025.html
Please cite as Bircan, T., A.A. Salah, "Emerging Big Data Sources for Studying Migration," in W. Allen, C. Vargas-Silva (eds.), Handbook of Research Methods in Migration, 2nd Ed., Edward Elgar Publishing, October 2024.

Keywords: Migration; big data; mobility indicators; big data ethics

Introduction

Over the last decades, global migration has been a major issue, along with the diffusion of technological applications that allowed the collection and linking of vast amounts of digital data. 'Migration preparedness' has been a focal point given humanitarian emergencies, such as in Syria, Ukraine, and Afghanistan, not to mention the consequences of global warming for human mobility, particularly in the Global South. Hyper-polarisation in the context of such social and political realms, refers to the increasing divergence of political attitudes to ideological extremes. In a hyper-polarised society, the discourse around migration often becomes highly charged and divisive, with different factions advocating for vastly different approaches to migration management. This can lead to a lack of consensus on how to prepare for and respond to migration emergencies, thereby complicating the planning and implementation of effective strategies. Moreover, hyper-polarisation can exacerbate the challenges associated with migration emergencies. For instance, it can fuel xenophobia and hostility towards migrants, further complicating efforts to manage migration in a humane and orderly manner. On the other hand, the urgency and scale of migration emergencies can also

contribute to hyper-polarisation. As societies grapple with the complexities of managing large-scale migration, differing views on how best to respond can become more entrenched, leading to increased polarisation.

Due to the complex and sensitive nature of the issue, the political rhetoric brings controversial terminologies together by conceptualising the 'preparedness' around *monitoring*, *predicting*, *crisis management*, and *border security*, while the new global[1] and European Union[2] migration pacts construe their mission and vision as better managed, regular, safe, and orderly migration. The controversy arises from differing perspectives on migration and its impacts, which are influenced by a range of factors including political ideology, socio-economic status, and personal experiences. For instance, some national policymakers view migration as a threat to national security, economic stability, or cultural identity, and therefore advocate for strict border controls and restrictive immigration policies. International alliances, however, underscore the fact that migration is a human right and an essential driver of economic growth and cultural diversity, and thus argue for more open and inclusive migration policies. These divergent views can lead to heated debates and polarisation, making it difficult to reach a consensus on migration management. This is particularly the case in the context of 'migration preparedness', where the need for effective strategies to manage large-scale migration emergencies often clashes with concerns about national sovereignty, security, and social cohesion. The controversy is not limited to any particular group but affects all stakeholders involved in migration issues, including policy-makers, migrants, host communities, and the general public. Understanding these controversies and their implications is crucial for developing balanced and effective migration policies and strategies. To achieve these ambitions, better data have been thought of as a key factor. Hence, in this context, the chapter contributes by exploring how big data can help navigate them to achieve better managed, regular, safe, and orderly migration.

Yet despite this potential, there is an ever-mounting body of literature that underlines the significant gaps in international migration statistics (Willekens et al., 2016; Bircan et al., 2020; Bosco, 2022; Salah et al., 2022a). These gaps are classified to be present in (1) definitions and measures, (2) migration drivers, (3) geographic coverage of the data, (4) migrants' demographic characteristics, and (5) timeliness of data acquisition (Ahmad-Yar and Bircan, 2021; Kraler and Reichel, 2022). It is now well established that the lack of sufficient, high-quality, and comparable migration statistics reduces the effectiveness of evidence-based policy approaches. Since the global-level migration indicators are relatively underdeveloped, there is a need to increase data and analytic capacities to generate timely, reliable, and comparable data on migration.

Efforts to build the knowledge base on migration-related flows, drivers, attitudes, and behaviours in quantitative terms are scattered across different disciplines. As a result, significant bonds have not been established yet. Creative data acquisition and analysis can

[1] https://www.iom.int/sites/g/files/tmzbdl486/files/our_work/ODG/GCM/NY_Declaration.pdf
[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0609

patch major shortcomings across the migration data landscape. With ever-expanding data storage and processing capabilities, new data sources are becoming available to study migration from additional perspectives, including, but not limited to social media data, mobile call data records, and satellite imaging. Except for a few studies assessing the digital data sources and data science applications for human migration and mobility (Sîrbu et al., 2021; Tjaden, 2021, Salah et al., 2022a), there remains a paucity of evidence on to what extent they can be worked towards an improvement of migration indicators. The aim of this chapter is to provide a systematic and concise review of potential big data sources that could provide valuable, real-time insights on migration that cannot be fully addressed through traditional data sources.

The remainder of this chapter proceeds as follows: The next section gives a brief overview of the long-standing migration measures based on traditional data sources. The third section is concerned with the use of big data sources for migration and mobility indicators. The fourth section amplifies big data analytics applications for assorted migration measures. After laying out the ethical considerations as major challenges of big data in the following section, the final part deals with the compilation of important insights into facilitating big data use for the development of migration measures. In total, this chapter aims to contribute to discussions around the promises and perils of big data for human migration research and monitoring.

1. Measures of Migration with Traditional Data

Migration studies is a pluralistic, cross- and inter-disciplinary field, which captures various aspects of human migration from conceptual, methodological, and empirical perspectives (Scholten et al., 2022). Disciplinary differentials in thematic and epistemological debates not only blur the scientific field boundaries (Hollifield, 2020) but also reflect on the adoption of varied empirical approaches in developing indicators for human migration. Before elaborating on the existing migration measures based on traditional data sources, we should describe how migration is quantified for and by policy frameworks.

The first classification is contingent on the geographical scope of the migration. Internal migration is concerned with people changing residences within a country (between cities, villages, etc.), and movements across international borders are considered as international migration.

While global policy categorisation of international human migration focuses mostly on the nature of migration (i.e., regular vs. irregular), scholarly work largely starts with the drivers of migration (i.e., voluntary vs. forced) (Findlay et al., 2015; Castles, 2020; Van Hear et al., 2020). However, international migration data collection is based on the former, and forced migration is implicitly covered by asylum application and decision data. Regular migration indicators cover documented/authorised entry and stay for the following reasons: (1) economic reasons, (2) educational reasons, (3) family (re)unification, and (4) humanitarian reasons. Recorded humanitarian reasons include forced migration, displacement and resettlement. Despite the above-mentioned gaps with the existing statistics, regular migration indicators are more

accessible and better documented when compared to irregular migration, where data and valid measures are very limited for a variety of reasons including deaths and disappearances, human trafficking, and smuggling.

Another essential concept for migration measures is the stock and flows approach, which is relevant for both regular and irregular migration. Migration stock is an indicator that represents the number of migrants in a given geography in a certain period of time. The starting restriction for migration is moving to another country for residential intentions for a long period. An international migrant defined by the United Nations (UN) is "a person who moves to a country other than that of his or her usual residence for a period of at least a year (12 months) so that the country of destination effectively becomes his or her new country of usual residence"[3]. International migration data sources (i.e., United Nations Department of Economic and Social Affairs - UNDESA, International Organization for Migration - IOM, Eurostat, Organisation for Economic Co-operation and Development - OECD, etc.) identify migrants based on 'country of nationality' and/or 'country of birth'. Although different and more detailed definitions can be used at the country and regional level by the national statistical offices, comparative data on international migration is harmonised mostly by nationality and country of birth. Migrant flows are composed of inflows (immigration) and outflows (emigration). The measurement of flows differs from stocks not only in terms of definition but also the geographical scope. Migrant stock indicators are produced for the host countries, whereas migrant flows can be estimated for the sending countries, for the host countries, or for modelling dyadic migration trends. Estimations for migration inflows are found to be more accurate given better data availability in receiving (Global North) countries (Ahmad-Yar and Bircan, 2021). Data on immigration flows include statistics on granted visas such as residence permits, work permits, study visas, family visas and statistics on granted asylum and protection.

The major aim of migration and mobility indicators is to provide an understanding of international human movements. Consequently, the migration statistics are predominantly built on migration stocks and flow estimations. Nevertheless, the living conditions of migrants in the host countries and their links with the sending countries are treated as integration indicators. Migrant integration measures can be produced based on data on remittances, residential and socio-economic segregation, access to public services, as well as inclusion and assessment of public opinion on migration and migrants.

Last, but not least, short-term human mobility is another domain where migration statistics are available. Existing statistics on short-term mobility cover cross-border mobility, tourism and commuting data. Even if the long-time and residential intention restrictions do not hold for this type of indicator, the temporal dynamics of these mobilities give rise to the challenges of keeping regular track of transient movements.

[3] UN Recommendations on Statistics of International Migration UN Department of Economic and Social Affairs, Statistics Division, Statistical Papers Series M No.No. 58, Rev.1, 1998.

Major traditional data sources for migration indicators are censuses, administrative sources, and (household) surveys. Censuses are the most comprehensive sources for producing demographic indicators including migration. However, they are mostly conducted every 5 or 10 years or have been replaced in some countries by population register data. Administrative sources are data register records and are used for creating statistics for administrative purposes. They can include residence information, health records, tax-related data, education statistics, social benefits, criminal records, police/border registers, visas, and work permits, etc. National household surveys and large-scale international surveys such as the World Values Survey, International Social Survey Programme (ISSP), Labour Force Survey (LFS), European Union Statistics on Income and Living Conditions (EU-SILC), European Quality of Life Surveys (EQLS), European Social Survey (ESS), etc. are crucial traditional data sources since they can deliver more detailed information on people (rather than registers/records) regarding diverse topics around socio-economic characteristics, attitudes, aspirations, access to services, social networks and experiences with migration.

In all three traditional data sources, the migrant populations in the country are identified by either nationality or country of birth. While in numerous countries, migrant descendants are also identified locally, data sources accessible from abroad provide information only for nationality and country of birth. Furthermore, inconsistencies between official definitions of certain statistics and their counterparts within the public and political rhetoric are discernible in numerous cases. The word 'migrant' can be used to denote migrant descendants, or 'refugee' and 'asylum seeker' might be used interchangeably, despite their prominent differences. To illustrate, a 'Ukrainian refugee' is not always a refugee but is typically recorded as a person 'under temporary protection' in official statistics. Similarly in Turkey, the millions of Syrians that came into the country after the war are referred to as 'Syrian refugees,' even though they are technically not refugees (due to a clause in the Geneva convention) and statistical records identify them as people 'under temporary protection'. Another challenge is that even though definitions focus on first-generation migrants, investigating migrant identity, integration, and societal acceptance of not only newcomers but also migrant descendants is a continuing concern in migration studies (Spencer and Charsley, 2021; Solano, 2022). Furthermore, intersectionality by various demographics, especially gender and education, is a major shortcoming of international migration data (Kofman, 2019; Bircan and Yilmaz, 2022). Moreover, these data sources include immigrants but not emigrants (Willekens et al., 2016) which leads to an incompatibility issue, particularly when migration statistics from sending and receiving countries are compared. Last but not least, time frames and references for official data collection/registration and the production of statistics create a challenge, which leads to significant time lags for data availability (UNECE, 2021). This issue concerns the temporal and geographical synchronisation of the registers and the statistics (e.g., emigrants might not deregister at the municipalities when leaving), but also the timeliness of the available most recent statistics (Ahmad-Yar and Bircan, 2021).

Despite suffering from varied shortcomings (Bircan et al., 2020), traditional data are and will be the major source for developing migration indicators. New data sources can be a potent

complement to traditional data for enhancing the existing measures and developing new indicators, where conventional data sources have identifiable issues.

2. Big Data Analytics for Migration and Mobility Indicators

Computational social science has investigated the potential of big data sources for analysing human behaviour at a large scale. The use of 'digital breadcrumbs' for migration research is a very recent development, and multiple big data sources offer different opportunities to develop proxies and indicators for migration. This offers the possibility to bridge the gap between official statistics and fast-moving migration trends (Sîrbu et al., 2021; Bosco et al., 2022; UNECE, 2022). The particular advantage of such sources is the unique insight they provide into simultaneous interactions and collective mobility in almost real-time and at a global scale, or with a granularity difficult to achieve with traditional methods. The specific challenge is factoring in the different kinds of biases incorporated in these sources. In this section, we will describe different sources and their uses in this domain.

3.1. Big Data Sources for Migration Research

The diffusion of new digital technologies in institutional and personal lives has become an accepted fact (for deeper insights, see Solove, 2004; Von Dijck, 2013; Schneier, 2015; Zuboff, 2023). Although public authorities and international organisations are still the leading official data producers, with increased activity in the digital life (such as the time spent on the internet, using smartphones, smart sensors, and automated algorithms), citizens have also become data producers through the digital traces they leave behind. Since the early definitions of Big Data in the mid-1990s (Cox and Ellsworth, 1997), traits that designate "Big Data" have also expanded. Laney (2001) argued that data can be identified as big data based on 3V's: Volume (size), velocity (speed) and variety (different types and structures). These traits require big data to consist of a vast amount of information that is created in (almost) real-time and can be presented in various structures. Kitchin and McArdle's (2016) seven axes approach adds exhaustivity (entire data), being fine-grained and uniquely indexical, relationality (having reference points for linkage), extensionality and scalability (being easily open for new fields and size expansion), veracity (being not error-free), value, and variability (context-dependent interpretations). To sum up, big data represents a different scale of analysis compared to what is previously accepted as the norm in social sciences and humanities. Big data typically refers to such a large quantity of information that it cannot be easily analysed by regular statistical programmes and require some computing power and programming language skills. It also represents a shift from precisely controlled subject pools to an assumption that when dealing with a huge number of subjects, some sources of variation will statistically get averaged out, and precise control is sacrificed in favour of coverage.

Regarding migration, three big data sources are prominently considered (Global Migration Group, 2017):

1. *Mobile-phone-based* - e.g., call records
2. *Internet-based* - e.g., social media

3. *Sensor-based* - e.g., earth observation data (satellite imagery).

In addition to these, other alternative data sources such as financial datasets, media archives, event datasets, etc. are also in use (Salah et al., 2022c). An increasing number of studies and initiatives are working on insights into migration phenomena that can be provided through the analyses of big data. We provide some examples to these sources and their usage here.

Mobile phone data: Mobile phones are carried on persons and communicate their approximate positions to a telecommunication operator. Consequently, they are (almost) ideal sensors to analyse mobility. Whenever a phone user makes a call or sends a text message, a line of information, including source and destination numbers, and connecting base towers (i.e., location) is stored by the telecom operator. These records (called call detail records - CDR), when appropriately anonymised and aggregated, can be used to study local movements of people and subpopulations. The number of calls at a given location is a strong proxy for how many people exist at that location at a given time. CDR data could also include time and duration of the call (which is useful to eliminate bot accounts making thousands of calls every day), anonymized statistics and demographics of the users, and anonymous identifiers used as proxies for the caller and the receiver. Because of the sensitivity of such data sets, ethical and legal guidelines are established for the aggregation, anonymisation, and sharing of mobile CDR data (Salah et al., 2019). Important shortcomings of this type of data are the difficulty of obtaining these records from a telecom operator, the biases of phone ownership (children, very poor, and very old may have inadequate representation), and the lack of insights about what causes changes (or anomalies) in the patterns. In countries where a phone operator has a large share of the market, the population coverage can be very high, and prior data initiatives have worked with data collected from millions of individuals.

Internet-based data: The Internet is not only the catalyst of big data generation, but it also enables previously impossible ways of data accumulation and collection, such as social media data, web browser search data, websites and application usage data (Blazquez and Domenech, 2018). Google Trends (GT) is an example of obtaining information from people's Internet search behaviour. It provides data on the volume of search queries for specific keywords and related themes allowing for longitudinal analyses going back to 2004. Social media is a significant source of information given the enormous number of active social media users globally, approximated as 4.7 billion in December 2022. 2.9 billion of those users were Facebook users, and TikTok reached 1 billion active users per month in 2022. Some platforms are particularly relevant for certain types of migration (e.g., LinkedIn for skilled migration). The popularity of these platforms, together with the geotagged information that can be extracted from them, can be leveraged to study mobility patterns (Vieira et al., 2022).

Sensor-based data: Over the last decades, remote sensing has become a valuable data source for performing large-area measurements of the physical conditions of the Earth's surface. Recent technological and methodological developments in satellite remote sensing have proven to provide highly detailed information on environmental conditions. Data from a variety of different sensors are available for large areas and free of cost, hence allowing researchers

and decision-makers to gather information and answer questions about environmental conditions and monitor global change. Sensor-based data include earth observation (satellite) and remote sensing (aerial imagery) data, as well as tracking data. These have been used for numerous cases related to human mobility including urbanisation, electrification, land use and vegetation, disaster monitoring, sea traffic, refugee settlements, etc. Satellites are orbiting objects equipped with earth observation sensors and cameras that collect imagery data from distance. There are 11,330 satellites orbiting the earth, 6,718 of which are active as of July 2023[4]. Most satellite data and related data products (e.g., environment and climate-related indices) are freely accessible (for a full list of satellite data sources, see Bircan, 2022). Tracking data include geolocation data, traffic and transport sensor data, closed-circuit television (CCTV) images, etc., and tracks from GPS embedded in smartphones, cars, and boats are used to model short-term human or sea vessel mobility (Pappalardo et al., 2015; Hashemi, 2017).

In terms of challenges, there is a growing concern regarding the political manipulation and misuse of these technologies for surveillance purposes. The use of unmanned aerial vehicles (UAVs) such as drones -mostly for military and security purposes- has expanded significantly and raised a considerable number of disputes, particularly for border management and surveillance. An important issue concerning the use of such data in real time relates to "push back operations", where refugee vessels are intercepted and sent back (Hayes and Vermeulen, 2012) and operating on historical data is less problematic.

Events data: Daily events occurring in the world are covered in multiple media channels, and repositories that store such media and allow indexing and retrieval into such archives present good opportunities of getting insights from events data. The Global Database of Events, Language and Tone (GDELT) Project[5] stores about 2.5 terabytes of media information per year, and monitors print, broadcast, and web news media across every country in the world, whereas the Armed Conflict Location & Event Data (ACLED) Project[6] accumulates information on political violence and protest events, with a similar global coverage. These repositories contain material from over 100 languages, and with the rapid progress in natural language processing (NLP) and machine learning (ML), sifting through these archives becomes easier. A much broader list of event datasets is provided by Ünver and Kurnaz (2022), who also discuss how social media can be harvested to create micro event datasets that rely more on field observations (of for instance violent events), rather than their potentially biased media coverage. These datasets provide important insights into what happened on a particular spot at a given moment, and can be combined with larger, but less detailed data sources. For instance, a mobile phone dataset may indicate an anomalous activity on a particular day in a specific location, and the event data may show that there was a political demonstration that probably caused the anomaly (Gündoğdu et al., 2016).

[4] https://www.unoosa.org/oosa/en/spaceobjectregister/index.html
[5] https://www.gdeltproject.org/
[6] https://acleddata.com/

Financial data: Financial transactions conducted by a person contain mobility and wealth related indicators, as well as behavioural signs of how much a person trades off exploration vs. exploitation in terms of resources nearby, providing insights into risk taking and plasticity (Singh et al., 2015). As such, datasets that record financial transactions can provide in-depth modelling opportunities for studying internal migration (Gürkan et al., 2022). These datasets, typically collected by banks and other financial institutions, can contain information about money withdrawals and deposits, money transfers, bill payments, credit card transactions and loan account activities. Of special interest to migration studies are remittances, which are money transfers from migrants to their home countries (Lim and Basnet, 2017) and financial anomaly detection for combating human trafficking (Moss, 2021).

3.2 Big Data for Migration-Related Estimations

Traditionally, official statistics are the major data source for developing estimates for migrant stocks, migration inflows/outflows, and other migration-related aspects such as short-term and urban mobility, migrant integration, and inclusion. By making innovative use of both traditional and big data sources and methods, several case studies have been conducted. As summarized in Table 1, we describe some important initiatives here, organised by the topics.

Table 1. A selection of big data sources for migration-related estimators

Estimation task	**New data sources (example works)**
Stocks	Facebook SCI (Alexander, 2021) Airtime top-up transfers (Aydoğdu et al., 2023; Bailey et al., 2018) Facebook Advertising Platform (Culora et al., 2021; Spyratos et al., 2019)
Flows	Facebook SCI (Goglia et al., 2022a; Minora et al., 2022) Twitter data (Hsiao et al., 2023) Satellite night lights (Chen, 2020)
--Short-term migration flows	Google trends (Avramescu and Wiśniowski, 2021; Fantazzini et al., 2021; Wanner, 2021)
--Asylum flows	Google trends and GDELT event data (Carammia et al., 2022)
--Skilled migration flows	LinkedIn data (Vieira et al., 2022)
--Internal migration	Financial transactions (Gürkan et al., 2022)
Short-term mobility	Satellite data (Bircan, 2022; Hashemi, 2017; Martin and Singh 2022; Momeni et al., 2021; Pappalardo et al., 2015) Mobile CDR (Beine et al., 2021; Tai et al., 2022) ACLED event data (Oishi et al., 2022) Social media data (Ünver, 2022)
Integration	Mobile CDR (Bakker et al., 2019; Hu et al., 2019; Salah et al., 2019) Mobile CDR and GDELT event data (Bertoli et al., 2019)

	Social media data (Arcila-Calderón et al., 2021; Kim et al., 2021; Pötzschke and Weiß, 2021)

New estimates for migration stocks

Official counts of migrant stocks can have data gaps for various reasons. Depending on the data collection period, official data may be outdated, and stocks coming from irregular migration can be poorly estimated. Novel data sources can provide more realistic estimates or raise flags about accuracy issues in the official data, but they need to be monitored for representativeness, as the services that generate such data come into and go out of fashion and corrected for such biases whenever possible.

An example is provided in Aydoğdu et al. (2023), where international airtime top-up transfer data were used as an indicator of migrant stocks. Such transfers happen typically when a migrant sends mobile phone credits to their home country. In their analysis, the authors have shown that if phone credit transfer service exists in a corridor, the distribution of migrant stocks in the country can be approximated by the amount of transfers from each region to the home country. Furthermore, corridors that show no officially registered migrants (such as United Arab Emirates - Uganda) may be revealed as having mobile transfer activity. The authors combine Facebook's Social Connectedness Index (SCI) with airtime top-up transfer activity to find 36 corridors with no official migrant stocks, but with high probability of having actual stocks. The difficulty in using air-time transfers is that the amount and frequency of transfers need not be strongly correlated with the number of stocks, and each corridor needs its own multiplier to establish a link between these variables. Furthermore, this type of data is difficult to obtain.

Facebook's SCI is an indicator of connectedness between two countries and is calculated as a normalised ratio of the number of friendship connections on Facebook between people from these countries (Bailey et al., 2018). This index is based on the Database of Global Administrative Areas (GADM v2.8) and the European Nomenclature of Territorial Units for Statistics (NUTS) areas. While the index is provided in a granular way at NUTS2 and NUTS3 levels, it excludes several countries where Facebook is banned or has few customers, or where there are political and legal reasons for not sharing such data (e.g., Afghanistan, Western Sahara, China, Cuba, Iraq, Israel, Iran, North Korea, Russia, Syria, Somalia, South Sudan, Sudan, Venezuela, Yemen). It has been used, jointly with other indicators, to predict migrant stocks in Goglia et al. (2022a), and to predict Ukrainian diaspora in Europe in Minora et al. (2022).

Another important source related to Facebook is their Advertising Platform, which provides group size estimates per area with demographic filters on age, sex, and country of origin through an Application Programming Interface (API). This data source was used by Alexander et al. (2020) to nowcast migrant stocks in the United States, and by others to estimate stocks

globally (Spyratos et al., 2019; Culora et al., 2021). To deal with systematic biases in Facebook related to demographics, researchers have used regression models to relate proportions of migrants in Facebook data to proportions in more controlled data collection efforts, such as the American Community Survey of the U.S. Census Bureau. Apart from biases, limitations of such data include the lack of transparency in the calculation of SCI, sudden changes in its estimation, missing values, and the influence of fake accounts (Culora et al., 2021).

Social media traces provide stock information from their users, and thus contain certain biases related to their user base. For example, LinkedIn is suitable for measuring stocks for skilled migrants (Vieira et al., 2022). If the particular selection biases for a platform are well known, it is possible to counter them via modelling (Alexander et al., 2020; Yildiz et al., 2017).

New estimates for migration flows

While most scholarly work focuses on estimating migration stocks as discussed in the previous section, there are noticeable efforts to use big data sources to develop new estimates altogether and to model migration flows. The pioneering approach using bilateral data for the analysis of international migration flows is gravity models to predict migration patterns (Ravenstein, 1885; Tinbergen, 1962; Anderson 2011). Analysing dyadic (country-pair) flows and stocks of international migration enables the identification of several critical factors, including the network effect, the influence of poverty constraints, and the impact of cultural links between countries. With the enhancement of alternative (big) data sources, dyadic analyses have been employed for migration flow estimates. We provide some examples of the adaptation of these fundamental concepts to big data analytics.

Goglia et al. (2022a) proposed a new measure, called the Bidirectional Migration Index (BMI), using inflows and outflows from two countries i and j for a specific year t:

$$BMI(t) = \frac{Flow_{i\to j}(t)+Flow_{j\to i}(t)}{Pop_i(t)\cdot Pop_j(t)}.$$

They utilised social media features, including Facebook's SCI, to develop regression models that predict BMI. In addition to the traditional features employed in gravity models, such as distance between countries, mean and difference in gross domestic product (GDP), and indicators for shared borders, religions, and languages, among others, their models included the SCI. The results demonstrated that incorporating SCI into the models enhanced their predictive ability, and the statistical analyses consistently indicated a strong and positive association with BMI.

Facebook as the one of the major social media platforms also provides various indices (in addition to SCI) and maps, through its Data for Good[7] programme. Another approach to use Facebook as a big data source is the Facebook advertising data. Alexander (2021) introduced a statistical framework for combining Facebook advertising data with traditional data sources

[7] https://dataforgood.facebook.com/dfg/tools

to estimate migrant stocks in North America through time series models. The study emphasised the importance of adjusting for non-representativeness (via post-stratification weighting in this specific case) to account for compositional biases, using historical demographic data, and accounting for errors in each data source.

Given the limitations of social media data, such as biased estimates due to lack of representativeness, Hsiao et al. (2023) proposed a method for combining Twitter data and official statistics by using the latter to model the spatial and temporal dependence structure of the biases of social media data. After confirming the validity of their approach through simulations based on the space-time framework, they estimated state-level out-migration probabilities in the United States by combining geo-located Twitter data with data from the American Community Survey (ACS). Their results indicated that combining (biased) social media and (unbiased) survey data produced more accurate predictions than relying solely on social media data. It is a notable illustration of how digital data can supplement, rather than substitute for, official statistics.

Satellite data can also serve as a potential big data source for predictive models for migration flows. An interesting empirical study is by Chen (2020), where night-time lights data was used as a proxy for human settlement and economic activities. This source was combined with total population and GDP per capita to estimate net migration flows (difference between inflow and outflow) in Europe. The study applied two types of statistical models (i.e., ordinary least squares regression and the spatial lag model) to analyse the relationship between Visible Infrared Imaging Radiometer Suite (VIIRS) lights and the net migration flows. The results demonstrated that night-time lights can be a useful tool for studying human migration patterns and may even be more effective than traditional measures like population size or GDP in certain countries.

Another example on estimating migration flows by combining traditional data with big data sources concentrated on asylum related flow predictions (Carammia et al., 2022). The predicted outcome variable was the weekly number of asylum applications (broken by nationality) lodged in each country member of the EU Common European Asylum System, and in the EU+. The study relied on big data to closely track migration drivers as they occurred, using event indices from the GDELT database and internet search data from Google Trends. Given the complexity of migration systems and drivers, separate models were developed for each individual country-to-country flow, and these models were trained using moving time windows to account for changes over time. This approach was applied to a broad range of bilateral asylum flows, originating from approximately 200 countries worldwide and directed towards each EU member state. Particular elaborative models were fit for 70 country-to-country flows, originating from seven countries of origin (Afghanistan, Eritrea, Iraq, Nigeria, Syria, Turkey, and Venezuela, respectively) and directed towards specific EU countries, as well as to the EU as a whole. The proposed model consistently outperformed benchmarks in almost all cases. Among the 43 dyads evaluated, the new (DynENet) model demonstrates strong performance in 30 instances, satisfactory results in four instances, subpar outcomes in three instances, and underperforms significantly in six instances. Essentially, within our varied sample, 70% of the

new forecasts are characterised as very good, while 80% are classified as either very good or good. Conversely, 20% of the forecasts yield poor or extremely poor results.

Another application with Google Trends data was proposed by Fantazzini et al. (2021), where the monthly migration data was combined with internet search volume data and other macro level regressors to predict internal migration flows in two largest cities of Russia (i.e., Moscow and St. Petersburg). The results of the traditional autoregressive integrated moving average (ARIMA) time series models showed that even after a series of robustness checks, incorporating Google Trends data into the models enhanced the accuracy of migration flow forecasting. Avramescu and Wiśniowski (2021) assessed the effectiveness of supplementing official migration statistics with Google Trends data to produce short-term forecasts of immigration flows from Romania to the United Kingdom. The composite variables that were created by Google Trends data (to show interest in migrating) were then analysed and compared to benchmark results using univariate time series models. Through a simpler model, Wanner (2021) tested the usefulness of Google Trends indices as a tool for predicting short-term migration flow to Switzerland from surrounding countries for the period of 2006-2016. The study demonstrated that the number of Google searches made by inhabitants of a country of departure can provide information on a country's attractiveness as a potential destination for migrants from that country. These studies illustrate that Google Trends indices may be useful for short-term forecasting of migration trends, but they also highlight that this should be used in conjunction with other data sources for more accurate and comprehensive long-term predictions. Furthermore, as people's search behaviour changes over time, prediction models require regular updates (Hofman et al., 2021).

New estimates for migrant integration

Local integration is seen as a legal, economic, and social process to ensure durable solutions for refugees (Crisp, 2004). New data sources provide indicators for these dimensions, provided that they can be aggregated separately for locals, and different migrant or refugee populations. For example, Hu et al. (2019) investigated Syrian refugees in Turkey via mobile CDR and point-of-interest data, and established that compared to locals, refugees had fewer high-expense activities, smaller travel distances, and different home locations. Bakker et al. (2019) looked at 1) the ratio of within group vs. between group calls for social integration, 2) the differential distribution of the minority population (i.e., evenness) and locations where minority and majority populations encountered each other (i.e., exposure) for spatial integration, and 3) commuting patterns (indicating regular work) for economic integration. Based on such indicators, it was possible to rank cities in terms of local integration of its refugee populations. In both cases, the primary data source was the Data for Refugees Challenge database, which provided mobile CDR disaggregated for locals and refugees (Salah et al., 2019), but additional data was used in conjunction with CDR. Bertoli et al. (2019) combined CDR with GDELT event data, as well as data on transactions on the housing market. The former is a good indicator of sentiments against the minority group, whereas the latter is a strong economic indicator.

As opposed to mobile CDR, which can have very high coverage and temporal resolution, social media data can be more sporadic, but provides richer insights into interactions and their polarity. Subsequently, social media data are used to gauge sentiments against minority groups and migrants (Arcila-Calderón et al., 2021), to measure segregation, and cultural integration (Kim et al., 2021). While social media can be used to boost the scale of traditional approaches, such as surveys (e.g., Pötzschke and Weiß, 2021), it can also provide content (e.g., text for sentiment analysis, links for social network analysis, image and video for content analysis) directly.

Big data sources typically provide aggregated information due to the difficulty of getting informed consent to process data at the individual level. Subsequently, they provide information at a higher level than traditional approaches to integration, and many complex interactions between the factors cannot be assessed.

New short-term mobility estimates

Big data sources can improve the understanding of the dynamics of environmental factors in migration. Satellites are an important data source for environmental indices, and their coverage can be exhaustive, as historical, and contemporary satellite imagery can provide monitoring of the entire globe (Bircan, 2022). One of the in-depth case studies using satellite data focused on Somalia (2016-2019) to understand the impact of different extreme weather events and climate factors on internal displacements (Momeni et al., 2021). Flood and drought indicators were developed through advanced GIS, image recognition and machine learning techniques. These indicators were combined with official statistics and register data to estimate internal displacements. The results showed that rapid-onset (i.e., flood) and slow-onset (i.e., drought) events have different effects on populations and satellite data could offer valuable insights to understand the link between the environmental factors and forced migration patterns.

Martin and Singh (2022) worked on two principal case studies, Somalia (2005–2007) and Iraq/Syria (2011–2017), to estimate population movements in the context of environmental change. In addition to official statistics and register data, they developed indicators from a new extensive dataset combining more than 700 million publicly available open-source media articles with news articles from around the world in 46 languages and more than 3 billion tweets in English and Arabic relevant hashtags, including those related to extremist groups. The study underlined the added value of big data in linking environmental factors to economic, social, demographic, and political drivers of human mobility, as the media data enabled gaining deeper insights into people's discussions, their viewpoints regarding various events and situations, and whether their comments and worries about local circumstances were on the rise or declined over time. With a similar interest on the conflict-forced migration, Oishi et al. (2022) forecasted internal displaced people movements in the Democratic Republic of the Congo with machine learning techniques on conflict data from ACLED, combined with geographical data including OpenStreetMap, Open Source Routing Machine, Territory Ethnic Composition, subnational population statistics and mining site information. The findings indicated that the predictive performance was enhanced as more open data was used for training the models.

Organised violence, along with extreme climate events, is a major cause of forced displacement. However, it is not easy to create predictive models, given the challenges with official data availability on violent events. The rise of social media use and development of event databases led scholars to investigate the use of social media data for migration and violence forecasting. Ünver (2022) focused on the debate regarding the use of such data to train machine learning classifiers in order to predict and monitor violence and forced migration using algorithmic systems. He identified and examined five common explanations in the literature for the limitations of using such data for AI forecasting, including policy-engineering mismatch, accessibility and comprehensibility, legal and legislative legitimacy, poor data cleaning, and difficulty of troubleshooting. The study suggested that solutions to these problems could include anonymisation, distributed responsibility, and the discussion of the "right to reasonable inferences".

Mobile CDR is suitable for analysing short-term mobility, particularly within a single country. In Beine et al. (2021), Syrian refugee movements within Turkey were investigated with CDR data from the Data for Refugees Challenge, and the influence of variations of income at origin and the distance were found to be important in determining mobility, as opposed to changes in income at destination. Tai et al. (2022) looked at Afghanistan (2013-2017) and using anonymised metadata from 10 million mobile phone users, detected home location changes across the country. Their results showed the distribution of duration of displacements following violence, as well as displacements that preceded (and anticipated) violence. Most importantly, this type of analysis provides insights about where the displaced are going, and the factors that play a role in displacement. In the aftermath of disasters and violence, reliable estimation of mobility is a crucial parameter for policy makers to allocate resources appropriately.

3. Challenges with Opportunities: Ethics, Big Data and Migration

The use of big data and new methodologies present significant opportunities to address the drawbacks in terms of aspects hard-to-estimate by official statistics, timeliness and to improve the migration governance systems. These methods also have certain drawbacks and difficulties, and ethical concerns emerge in the field of migration and mobility data science, stemming from the sensitivity of migration as a domain and the potential (mis)use of innovative technological tools related to migration (Molnar, 2019; Salah et al., 2022b).

While researchers in the field have increasingly faced the paradoxical problem of data abundance, the use of big data for migration analysis raises ethical questions about the potential impact of the analysis on the individuals and communities that are the subject of the analysis, as well as broader questions about the ethical use of data and technology in society. A major issue is the clash between the interests of different stakeholders, as the same technologies can be used for assisting migrants as well as initiating pushbacks. But even if we assume benevolent institutions all around, there are several concerns about this data upsurge.

The first issue is the challenge of working with partial, imperfect, and biased data. The accuracy and completeness of big data used in migration analysis can be problematic, particularly in contexts where gaps in migration statistics are significant (Ahmad Yar and Bircan, 2021). Newly developed indicators and models require robustness checks with the use of official records. In the cases where grassroots data is lacking, new estimates and models cannot be assessed in terms of validity and reliability. There are sampling and selection biases (non-representativeness of the big data source), algorithmic biases (shortcomings of the specific algorithms), and availability biases (working with "found" data), and the potential issues of accuracy and quality raise concerns about trustworthiness and about the consequences of basing policy decisions on shaky ground. As the analysis of big data relies more and more on artificial intelligence (AI) techniques, it inherits problems of such systems. AI algorithms that learn from human behaviour and data end up incorporating real life inequalities, socio-political and cultural biases, as well as past discrimination or injustices, thus perpetuating inequalities in the present. The interpretation and analysis of big data also requires a high level of expertise and knowledge, particularly in relation to the context in which the data was collected and the potential biases that may be present in the data. For migration related applications, contextual dynamics, human rights aspects, and vulnerabilities should be specifically well-acknowledged to understand the real-life implications of the analyses' results.

Secondly, sifting through data is a much bigger task than it was in the past. This warrants larger resources as well as empowered data collaborations with public institutes and the private sector not only big tech companies but also SMEs. *Data collaboratives* are public-private partnerships based on data owned by private parties. Different models have been proposed for such collaborations, but the main idea is that private data is re-purposed for the public good. *Data challenges* are a form of data collaboratives, where the private data are processed to remove personal information and opened to a larger research community, allowing multiple groups to analyse it from different aspects (Salah et al., 2019). While there are hardly any collaboratives sustained over long periods yet, there are initiatives from both governments and industry to find ways of enabling them. The requirements include establishing proper legal frameworks to enable data exchange, developing technological solutions for timely and ethical data and insight sharing, and overcoming social hurdles related to trust. The collection and use of sensitive personal data can raise privacy and confidentiality concerns, particularly in contexts where data protection laws may be weak or non-existent. The ownership of data used in migration analysis can be unclear. This obfuscates determining who is responsible for the accuracy and appropriateness of the analysis, particularly considering institutional interests of political actors which shape decisions about data collection (Scheel and Ustek-Spilda, 2018).

Thirdly, but not less important, scholars and civil society have expressed profound concerns about the emerging reliance on AI-powered technologies for managing border spaces, automatising administrative decision making and regulating migratory movements. These concerns relate to the use of technology as a solution to the intricate challenge of data use, data access related to migration governance. AI technologies and the databases underpinning them can pose a risk to migrants' right to privacy and hence to the protection of migrants (Beduschi, 2018). While inadequate regulation renders migrants more trackable and intelligible (Molnar,

2019), lacking an effective auditing system induces accountability obscurity among states, international institutes, and corporations (Bircan and Korkmaz, 2021). The legislative frameworks, confidentiality and engagement practices lie at the heart of ethical concerns about human rights violations through the use of big data and imposing pseudo and biased 'evidence' to introduce more conservative and preventive migration policies. Despite all challenges, harnessing the opportunities of these new data sources and innovative methodologies can only be realised with the active involvement of social scientists and citizen (migrant) engagement.

4. Conclusions and Future Directions

This chapter highlights the importance of a need for a versatile and a comprehensive approach to understand the complex and dynamic phenomenon of migration through a multi-disciplinary lens. The use of big data sources with advanced analytics can provide valuable insights with high spatio-temporal granularity, although more coordinated and significant efforts are needed to enhance the knowledge base on migration-related challenges in quantitative terms via new data sources (Salah et al., 2022a).

Big data sources and analytics have become increasingly popular tools in social science research to develop new indicators and complex models for reliable estimations and automated systems, but it is important to note that these should not be seen as a replacement for traditional data sources. Instead, big data can be used to complement traditional data and fill in gaps in research, to enhance explanatory modelling with prediction capabilities (Hofman et al., 2021). To fully utilise big data in social science research, social scientists need to develop further technical skills and collaborate with computer and data scientists to foster interdisciplinary research. The results of predictive models, analysis or systems built on big data may be influenced by biases that may yield outcomes lacking transparency and accountability. Therefore, before these applications are initiated for practical use, the related risks must be openly addressed and resolved.

Last but not least, there are potential societal risks associated with utilising big data sources to track human mobility and migration. Ethical and socially responsible use of big data is crucial, especially when it comes to migration policies. It is important to ensure that big data is not used to grant unwarranted surveillance capabilities to governments and corporations and hence to serve restrictive and predictive migration policies that can have negative impacts on marginalised communities.

While we did not describe individual AI technologies used in the processing of big data sources, they are essential in forming complex connections between big data sources and target variables such as indicators. As the amount of data increases, the potential of models to capture influences between factors increases proportionally. As the AI models grow in complexity, we see large-scale efforts to improve their transparency and explainability, so that they can be employed for assisting humans better in decision making. Ultimately, AI is perceived to be a means to create automation, and to reduce human costs across a wide range of applications. For migration indicators, costly census studies and time-consuming interviews cannot be fully

replaced by big data-based solutions but made cheaper and more targeted if these approaches are combined.

In addition to improving transparency of AI approaches, a second important trend is in the regulations imposed on the use of AI. By making sensitive data harder to harvest, we protect people from potential misuse of such data, but also make the training of AI algorithms harder. Different regulations in different countries will doubtless affect the development of local solutions, just as different data practices affect collection and processing of migration related indicators now. With the automated tools, however, more bias may enter the analyses, as tools developed in one context may not fit well in other contexts.

Lastly, while AI is making some tasks cheaper through automation, it introduces two important problems. First of all, it removes the flexibility and quality of expert human solutions. The obvious counterargument is that such expert resources are not uniformly available. Secondly, while we may assume that some tasks are too sensitive to be automated, the temptation may prove to be too great. As specific failure cases (such as the case of using AI in prediction of recidivism, see Dressel and Farid, 2018) come to light, they serve as important beacons for the moderation and re-assessment of AI and big data-based solutions.

Acknowledgement

This work was supported by the European Commission through the Horizon2020 European project “HumMingBird: Enhanced migration measures from a multidimensional perspective” (Research and Innovation Action, grant agreement no 870661).

References

Ahmad-Yar, A. W., and Bircan, T. (2021). Anatomy of a Misfit: International Migration Statistics. *Sustainability*, 13(7), 4032.

Alexander, M. (2021). Using social media advertising data to estimate migration trends over time. In: *Big Data Applications in Geography and Planning* (pp. 8-24). Edward Elgar Publishing.

Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the United States. *Population Research and Policy Review*, 1-28.

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics*. 3(1), 133-160.

Arcila-Calderón, C., Blanco-Herrero, D., Frías-Vázquez, M., and Seoane-Pérez, F. (2021). Refugees welcome? Online hate speech and sentiments in Twitter in Spain during the reception of the boat Aquarius. *Sustainability*, 13(5), 2728.

Avramescu, A., and Wiśniowski, A. (2021). Now-casting Romanian migration into the United Kingdom by using Google Search engine data. Demographic Research, 45, 1219-1254.

Aydoğdu, B., Samad, H., Bai, S., Abboud, S., Gorantis, I. and Salah, A.A. (2023), Analyzing international airtime top-up transfers for migration and mobility. *International Journal of Data Science and Analytics,* https://doi.org/10.1007/s41060-023-00396-7.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., and Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3), 259-80.

Bakker, M. A., Piracha, D. A., Lu, P. J., Bejgo, K., Bahrami, M., Leng, Y., ... and Pentland, A. (2019). Measuring fine-grained multidimensional integration using mobile phone metadata: the case of Syrian refugees in Turkey. In: Salah, A.A., Pentland, A., Lepri, B., Letouzé, E. (Eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, 123-140, Springer: Cham.

Beduschi, A. (2018). The Big Data of International Migration: Opportunities and Challenges for States under International Human Rights Law. *Georgetown Journal of International Law,* 49/2: 982–1017.

Beine, M., Bertinelli, L., Cömertpay, R., Litina, A., and Maystadt, J. F. (2021). A gravity analysis of refugee mobility using mobile phone data. *Journal of Development Economics,* 150, 102618.

Bertoli, S., Cintia, P., Giannotti, F., Madinier, E., Ozden, C., Packard, M., ... and Speciale, B. (2019). Integration of Syrian refugees: insights from D4R, media events and housing market data. In: Salah, A.A., Pentland, A., Lepri, B., Letouzé, E. (Eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, 179-199, Springer: Cham.

Bircan, T. (2022). Remote Sensing Data for Migration Research. In: Salah, A.A., Korkmaz, E.E. and Bircan, T. (Eds.) *Data Science for Migration and Mobility.* Proceedings of the British Academy. London: British Academy / Oxford University Press.

Bircan, T., and Korkmaz, E. E. (2021). Big Data for Whose Sake? Governing Migration through Artificial Intelligence. *Nature Humanities and Social Sciences Communications*, 8(1), 1-5.

Bircan, T., Purkayastha, D., Ahmad-Yar, A.W., Lotter, K., Iakono, C.D., Göler, D. et al. (2020) Gaps in Migration Research. Review of Migration Theories and the Quality and Compatibility of Migration Data on the National and International Level, *Deliverable 2.1 of the HumMingBird project*. Available from: https://hummingbird-h2020.eu/images/publicationpdf/d2-1-eind-1.pdf [Accessed on 29 November 2022].

Blazquez, D., and Domenech, J. (2018). Big Data Sources and Methods for Social and Economic Analyses. *Technological Forecasting and Social Change*, 130, 99-113.

Bosco, C., Grubanov-Boskovic, S., Iacus, S., Minora, U., Sermi, F., and Spyratos, S. (2022). Data Innovation in Demography, Migration and Human Mobility. *arXiv preprint arXiv:2209.05460*.

Carammia, M., S. M. Iacus, and T. Wilkin (2022). Forecasting asylum-related migration flows with machine learning and data at scale. Nature Scientific Reports (SREP-21-01036)

Castles, S., 2000. International migration at the beginning of the twenty-first century: Global trends and issues. *International Social Science Journal*, 52(165), pp.269-281.Chen, X. (2020). Nighttime lights and population migration: Revisiting classic demographic perspectives with an analysis of recent European data. Remote Sensing, 12(1), 169.

Cox, M., and Ellsworth, D. (1997). Managing Big Data for Scientific Visualization. *ACM Siggraph*, 97(1), 21-38.

Crisp, J. (2004). The local integration and local settlement of refugees: a conceptual and historical analysis (pp. 1-8). Geneva, Switzerland: UNHCR.

Culora, A., Thomas, E., Dufresne, E., Cefalu, M., Fays, C., and Hoorens, S. (2021). Using Social Media Data to 'Nowcast' International Migration Around the Globe. RAND.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1), p.eaao5580.

Fantazzini, D., Pushchelenko, J., Mironenkov, A., and Kurbatskii, A. (2021). Forecasting Internal Migration in Russia Using Google Trends: Evidence from Moscow and Saint Petersburg. Forecasting, 3(4), 774-803.

Findlay, A., McCollum, D., Coulter, R. and Gayle, V., 2015. New mobilities across the life course: A framework for analysing demographically linked drivers of migration. *Population, Space and place*, 21(4), pp.390-402.

Goglia, D., Pollacci, L., and Sîrbu, A. (2022a). Dataset of Multi-aspect Integrated Migration Indicators. *arXiv preprint arXiv:2204*.14223.

Goglia, D., Pollacci, L., and Sîrbu, A. (2022b). Use of non-traditional data sources to nowcast migration trends through Artificial Intelligence technologies. In *Measuring Migration Conference 2022 Conference Proceedings* (Vol. 42, p. 15). Transnational Press London.

Gündoğdu, D., Incel, O. D., Salah, A. A., and Lepri, B. (2016). Countrywide arrhythmia: emergency event detection using mobile phone data. *EPJ Data Science*, 5(1), 25.

Gürkan, M., Bozkaya, B. and Balcısoy, B. (2022). Financial datasets: Leveraging transactional big data in mobility and migration studies. In: Salah, A.A., Korkmaz, E.E. and Bircan, T. (Eds.) *Data Science for Migration and Mobility.* Proceedings of the British Academy. London: British Academy / Oxford University Press.

Hashemi, M. (2017). Intelligent GPS trace management for human mobility pattern detection. *Cogent Engineering*, 4(1), 1390813.

Hayes, B., and Vermeulen, M. (2012). *The EU's New Border Surveillance Initiatives.* Berlin: Heinrich Böll Foundation.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... and Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866), 181-188.

Hollifield, J. F. (2020). Is Migration a Unique Field of Study in Social Sciences? A Response to Levy, Pisarevskaya, and Scholten. *Comparative Migration Studies*, 8(1), 1-9.

Hsiao, Y., Fiorio, L., Wakefield, J., and Zagheni, E. (2023). Modeling the Bias of Digital Data: An Approach to Combining Digital With Official Statistics to Estimate and Predict Migration Trends. Sociological Methods and Research, 00491241221140144.

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

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

Kitchin, R., and McArdle, G. (2016). What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data & Society,* 3(1).

Kofman, E. (2019). Gendered Mobilities and Vulnerabilities: Refugee Journeys to and in Europe. *Journal of Ethnic and Migration Studies*, 45(12), 2185-2199.

Kraler, A., and Reichel, D. (2022). Migration Statistics. In Scholten, P. (Ed). *Introduction to Migration Studies An Interactive Guide to the Literatures on Migration and Diversity*. Springer, Cham, 439-462.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note*, 6(70), 1.

Lim, S. and Basnet, H. C. (2017). International migration, workers' remittances and permanent income hypothesis, *World Development,* 96, 438-450.

Martin, S. F., and Singh, L. (2022). Environmental change and human mobility: Opportunities and challenges of big data. *International Migration*.

Molnar, P. (2019). Technology on the Margins: AI and Global Migration Management from a Human Rights Perspective. *Cambridge International Law Journal,* 305–30.

Momeni, R., Bircan, T., King, R. (2022). Environment-induced internal displacement. The case of Somalia (Deliverable 4.3). Leuven: HumMingBird project 870661 – H2020.

Moss, I. E. K. (2021). *Role of Forensic Accounting in Combating Global Human Trafficking*. Doctoral dissertation, Northcentral University.

Oishi, A., Teshima, T., Akao, K., Kano, T., Kiriha, M., Kojima, N., ... and Yamanaka, S. (2022). Forecasting Internally Displaced People's Movements with Artificial Intelligence. In *Digital Innovations, Business and Society in Africa,* . Springer, Cham, pp.311-339.

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., and Barabási, A. L. (2015). Returners and Explorers Dichotomy in Human Mobility. *Nature Communications*, 6(1), 1-8.

Pötzschke, S. and Weiß, B. (2021). Realizing a Global Survey of Emigrants Through Facebook and Instagram. OSF Preprints. October 11. doi:10.31219/osf.io/y36vr.

Ravenstein, E. G. (1885). *The laws of migration.* Royal Statistical Society.

Salah, A.A., Bircan, T. and Korkmaz, E.E. (2022a). New Data Sources and Computational Approaches on Migration and Human Mobility. In: Salah, A.A., Korkmaz, E.E. and Bircan, T. (Eds.) *Data Science for Migration and Mobility.* Proceedings of the British Academy. London: British Academy / Oxford University Press.

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

Salah, A.A., Korkmaz, E.E. and Bircan, T. (2022c). *Data Science for Migration and Mobility.* Proceedings of the British Academy. London: British Academy / Oxford University Press.

Salah, A. A., Pentland, A., Lepri, B., Letouzé, E., De Montjoye, Y. A., and Vinck, P. (2019). *Guide to mobile data analytics in refugee scenarios*. Cham: Springer.

Scheel, S., and Ustek-Spilda, F. (2018). Big data, big promises: Revisiting migration statistics in context of the datafication of everything. *Border Criminologies.*

Scholten, P., Pisarevskaya, A., and Levy, N. (2022). An Introduction to Migration Studies: The Rise and Coming of Age of a Research Field. In *Introduction to Migration Studies*. Springer, Cham. 3-24.

Schneier, B., 2015. *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company.

Singh, V. K., Bozkaya, B. and Pentland, A. (2015). Money walks: Implicit mobility behavior and financial well-being. *PloS One* 10(8), 1-17.

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., ... and Sharma, R. (2021). Human Migration: The Big Data Perspective. *International Journal of Data Science and Analytics*, 11(4), 341-360.

Solano, G. (2022). Indicators and Survey Data to Understand Migration and Integration Policy Frameworks and Trends in the EU. In: Salah, A.A., Korkmaz, E.E. and Bircan, T. (Eds.) *Data Science for Migration and Mobility.* Proceedings of the British Academy. London: British Academy / Oxford University Press.

Solove, D.J., 2004. *The digital person: Technology and privacy in the information age* (Vol. 1). NyU Press.

Spencer, S., and Charsley, K. (2021). Reframing 'Integration': Acknowledging and addressing five core critiques. *Comparative Migration Studies*, 9(1), 1-22.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. (2019). Quantifying international human mobility patterns using Facebook Network data. *PloS One*, 14(10), e0224134.

Tai, X. H., Mehra, S., and Blumenstock, J. E. (2022). Mobile phone data reveal the effects of violence on internal displacement in Afghanistan. *Nature human behaviour*, 6(5), 624-634.

Tinbergen J. (1962). *Shaping the World Economy: Suggestions for an International Economic Policy.* New York: Twentieth-Century Fund.

Tjaden, J. (2021). Measuring migration 2.0: a review of digital data sources. *Comparative Migration Studies,* 9(1), 1-20.

UNECE (2021). Guidance on the Use of Longitudinal Data for Migration Statistics. *ECE/CES/STAT/2020/6*.

UNECE (2022). Use of New Data Sources for Measuring International Migration. *ECE/CES/STAT/2022/7*.

Ünver, H. A. (2022). Using social media to monitor conflict-related migration: A review of implications for AI Forecasting. *Social Sciences*, *11*(9), 395.

Ünver, H.A. and Kurnaz, A. (2022). Conflict and forced migration: Social media as event data. In: Salah, A.A., Korkmaz, E.E. and Bircan, T. (Eds.) *Data Science for Migration and Mobility.* Proceedings of the British Academy. London: British Academy / Oxford University Press.

Van Dijck, J., 2013. *The culture of connectivity: A critical history of social media*. Oxford University Press.

Van Hear, N., Bakewell, O. and Long, K., 2020. Push-pull plus: reconsidering the drivers of migration. In *Aspiration, Desire and the Drivers of Migration* (pp. 19-36). Routledge.

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

Yildiz, D., Munson, J., Vitali, A., Tinati, R., and Holland, J. A. (2017). Using twitter data for demographic research. *Demographic Research*, 37, 1477–1514.

Wanner, P. (2021). How well can we estimate immigration trends using Google data?. Quality & Quantity, 55(4), 1181-1202.

Willekens, F., Massey, D., Raymer, J., and Beauchemin, C. (2016). International Migration under the Microscope. *Science*, 352(6288), 897-899.

Zuboff, S., 2023. The age of surveillance capitalism. In *Social Theory Re-Wired* (pp. 203-213). Routledge.