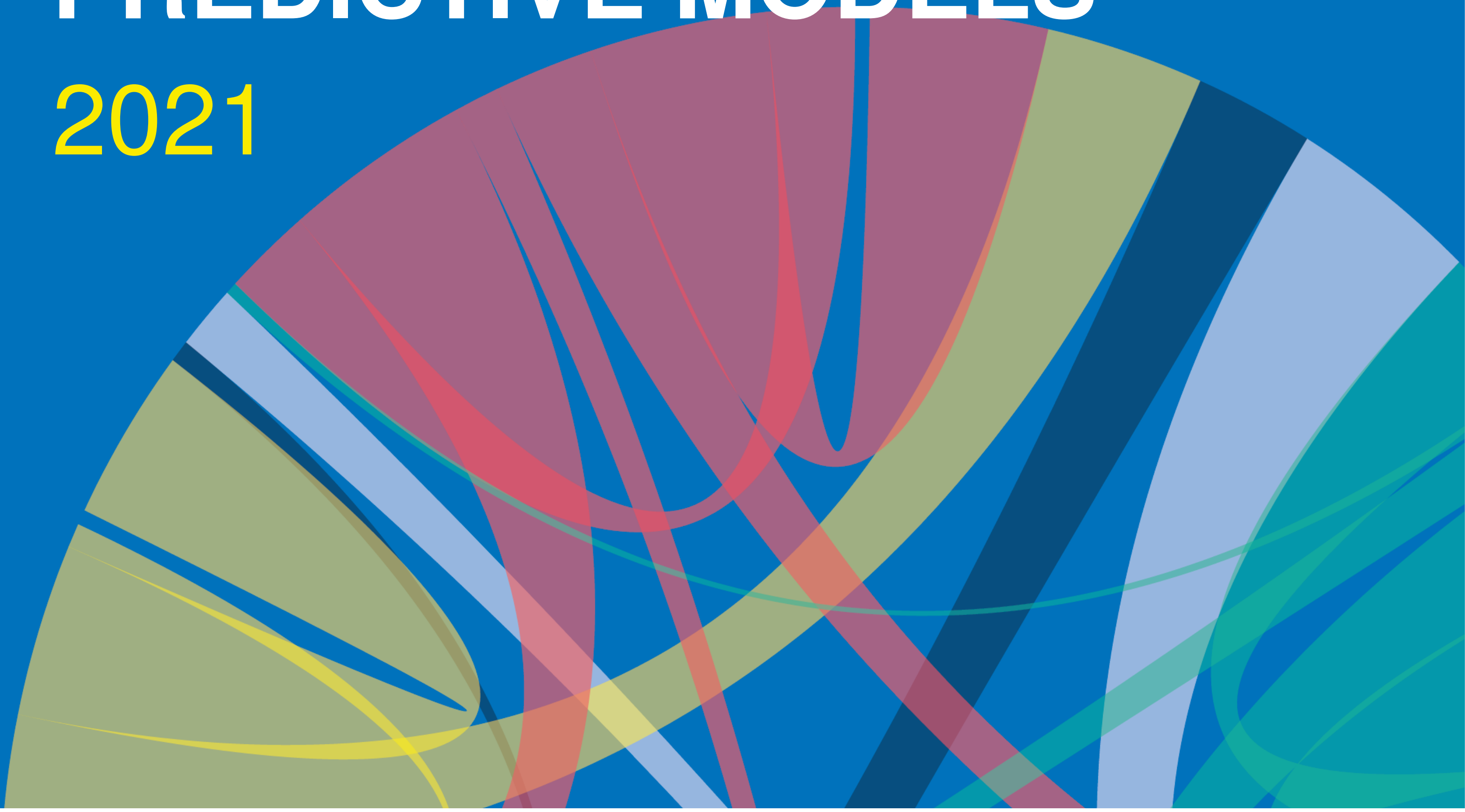
UNHCR
The UN Refugee Agency

BIG (Crisis) DATA FOR PREDICTIVE MODELS

2021



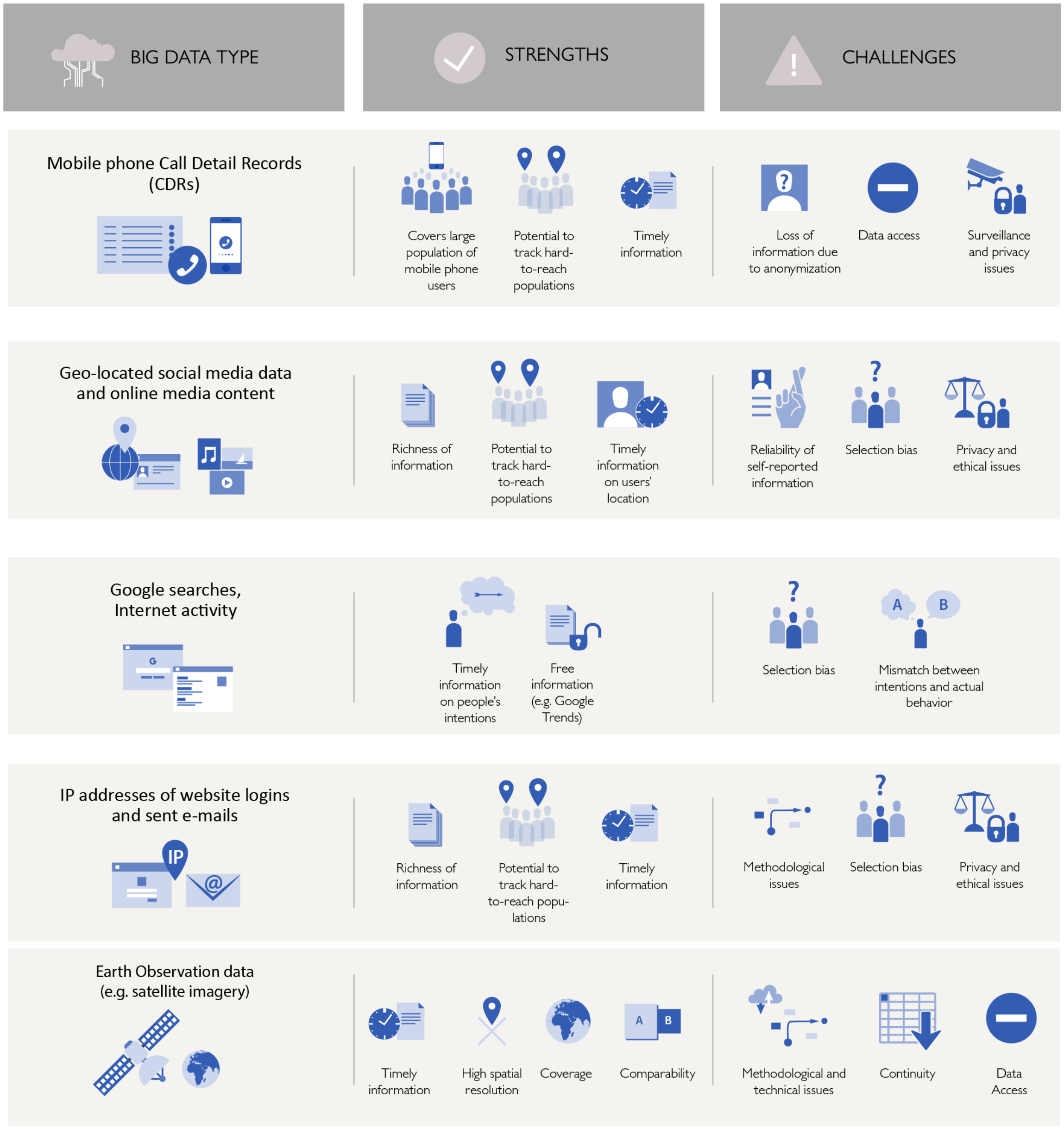
A Literature Review

Published by Statistics and Demographics Section
UNHCR Global Data Service
Contact Geraldine Henningsen, stats@unhcr.org

Overview of Sources of Big (Crisis) Data

The potential of big data and innovative data sources

Filling the gaps of traditional data sources on migration (censuses, surveys and administrative sources) and contributing to improved understanding of various aspects of migration. A number of studies have shown such potential – but there are significant challenges too.



BIG DATA TYPE	STRENGTHS	CHALLENGES
Mobile phone Call Detail Records (CDRs)	Covers large population of mobile phone users Potential to track hard-to-reach populations Timely information	Loss of information due to anonymization Data access Surveillance and privacy issues
Geo-located social media data and online media content	Richness of information Potential to track hard-to-reach populations Timely information on users' location	Reliability of self-reported information Selection bias Privacy and ethical issues
Google searches, Internet activity	Timely information on people's intentions Free information (e.g. Google Trends)	Selection bias Mismatch between intentions and actual behavior
IP addresses of website logins and sent e-mails	Richness of information Potential to track hard-to-reach populations Timely information	Methodological issues Selection bias Privacy and ethical issues
Earth Observation data (e.g. satellite imagery)	Timely information High spatial resolution Coverage Comparability	Methodological and technical issues Continuity Data Access

© IOM's GMDAC 2017 www.migrationdataportal.org

Satellite Imagery: Satellites, and to a lesser extent unmanned aerial vehicles (UAV) (e.g., drones), produce day-time imagery, radar measures, or nocturnal light images within more or less regular time intervals and with different granularity. In a humanitarian context, the granularity of images from older satellites is sufficient to analyse night-time lightning, whilst modern satellites can produce images with high resolution that allow identifying even small structures, like, e.g., tents, debris, or groups of people. The movement of people on the ground can, in some cases, be assessed by a rapid sequence of images taken by the satellite.

Cloudiness can impact day-time and night-time satellite imagery and create gaps in the data when it persists over extended periods. also, satellites will not cover all areas of the globe at the same frequency, which means that data on certain areas will be sparse. Radar imagery is not affected by clouds, as are UAVs which, on the other hand, only have a limited range.

Programmes such as Global Pulse's PulseSatellite programme in collaboration with UNOSAT or the Joint Research Center of the European Commission's GHSL programme, or NASA provide satellite imagery for humanitarian work. Otherwise, low resolution satellite imagery is freely available from Landsat or Copernicus, whilst high resolution satellite imagery is available from commercial providers, e.g., Digital Globe, Airbus, or Planet.

News Articles and Blogs: News articles from all major news agencies and blog posts are daily collected and categorised through detailed event keys in databases. The currently biggest databases are EventRegistry, GDELT, which is publicly available, and ICEWS, with only limited access. Further smaller projects include, e.g., Georgetown University's EOS News Article Archive or GTD the Global Terrorism Database provided by the University of Maryland.

Social Networks: Social network sites like Facebook, Instagram, Twitter, or Tumblr provide a constant flow of user-generated content, e.g., posts, pictures, microblogs. Some providers allow to geotag the content, which enables the analyst to connect user-generated content to geographical areas; however, studies show that geotagging accuracy is limited and often deactivated on purpose. Most social media sites allow limited access to their data via a public API (see table 1 in the appendix for an overview). However, the largest (in terms of number of users) social network site Facebook only offers pre-processed, aggregated user data via their advertising platform or Facebook Connect, which both aggregate Facebook's user data to demographic indicators, e.g., users' ex-pat status. Facebook's 'Data for Good' initiative offers several pre-processed products for humanitarian organisations on the basis of bilateral access agreements. Also, Twitter signed an agreement with UN Global Pulse which gives access to the Decahose and allowed UN Global Pulse to build the AI text processing tool Qatalog.

Call Detail Records: Call detail records (CDR) are user data collected for billing purposes by carrier networks. Every time a user calls, texts, or accesses the internet through her phone, the cell phone connects to a cell tower which allows approximating the user's location. Location tracking through CDRs is one of the most accurate methods to track population movements. However, the data are proprietary to the carrier network and cannot be publicly accessed. Analysts can access CDRs from a limited number of networks through Flowminder.

IP-addresses: Websites that require regular log-ins, e.g., websites of e-mail providers, store the user's IP-addresses. As these IP-addresses are connected to a user account, the user's movements can be tracked through the changes in IP-addresses from multiple log-ins. The data are proprietary to the service provider and cannot be publicly accessed.

Search Histories: Google Trends allows extracting aggregated search histories filtered by keyword and user's IP-address location. The keyword searches are indexed on a 0 to 100 scale, which is the relative share of the keyword search given all inquiries within a given region and time frame. The indexed keyword searches enable the analyst to generate standardised time series which reflect the popularity of a specific keyword over time. The data are publicly available.

Other sources: Other Big Data sources that have potential interest in the prediction of forced displacement are data that are regularly posted online, e.g., climatic and weather data, price data from commodity exchanges, or price data from local food markets published by the FAO, national governments, or NGOs. A further potential source of Big (Crisis) Data are crowd-sourced crisis maps, e.g., Ushahidi, HOTOSM, or Liveuamap.

Introduction

The recent Global Compact for Refugees has acknowledged that the increasing number of forcibly displaced persons and the difficulty to predict mass-movements of people have created an urgent need for data-driven early warning systems that allow governments and humanitarian organisations to use their limited resources most efficiently [1, 2]. Efforts to install early warning systems have led to advances in predicting and forecasting global migration flows; however, forced displacement remains the most elusive and challenging migration form to predict [3, 2, 4].

People base their decision to migrate on a complex set of different factors. However, in addition to the complexity of the individual decision-making process, predicting forced displacement flows is further complicated by the necessity to early detect or predict trigger events. Trigger events, which are often the last link in a long chain of other events that tilt the individual's decision towards flight, often happen randomly and abruptly [5]. Both processes are difficult to model in and by themselves. However, the lack of timely and accurate data at the micro-, meso-, and macro-level further aggravates this problem. Data needed to model these processes are often either unavailable (micro-level) or outdated (meso- and macro-level).

Furthermore, whilst existing refugee flows between countries often are self-perpetuating and can be predicted based on historical data, refugee flows from new events pose a challenge because (i) the event might no yet be known to the modeler; (ii) the effect of a new event on future refugee flows is unknown; (iii) no historical data from a recent event exist, and it is uncertain to which degree historical data from other events are applicable to predict refugee flows from the new event. The prediction of forced displacement flows requires, therefore, first and foremost, a thorough understanding of the mechanisms of forced displacement. In particular:

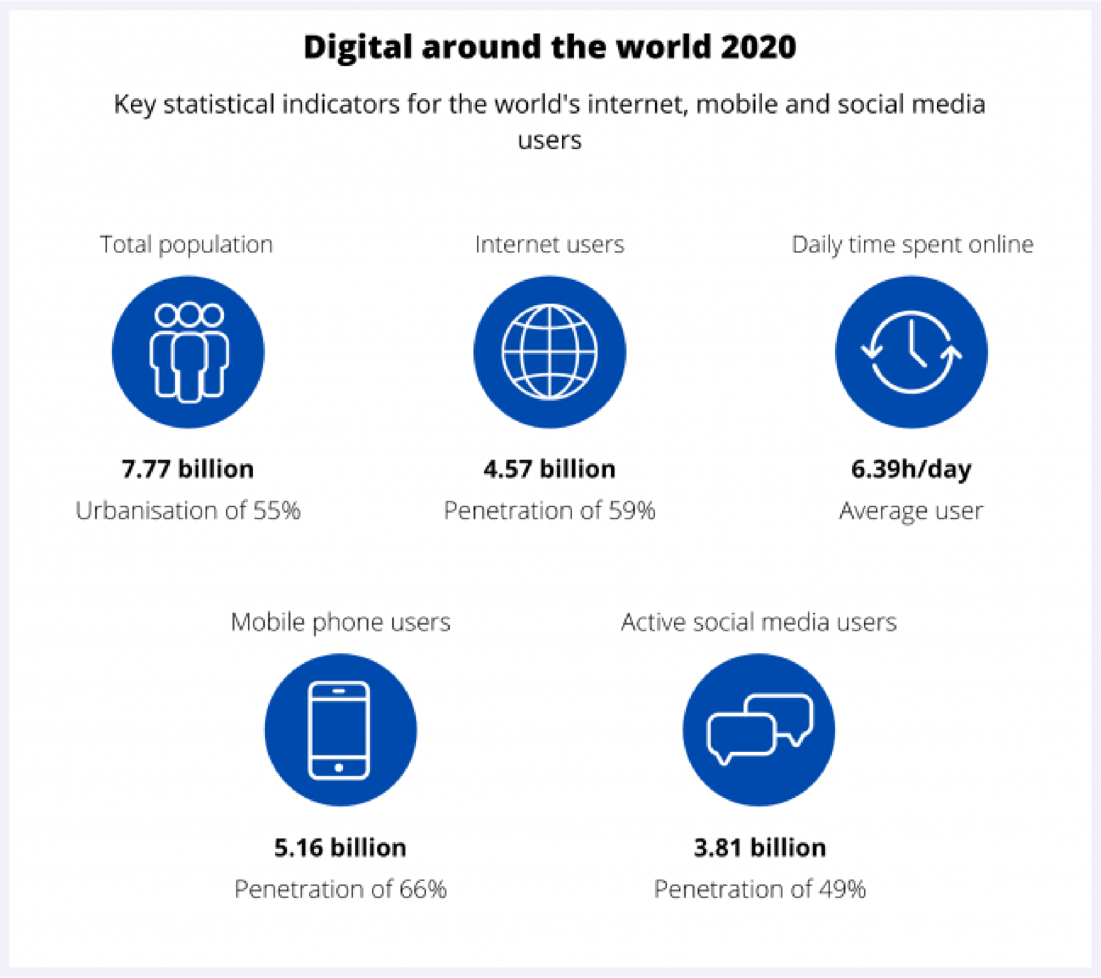
1. The factors that lead to an event that potentially can trigger forced displacement.
2. The characteristics of an event that have the potency to create sizable forced displacement.
3. The factors that impact the magnitude, demographic, and direction of forced displacement.

Furthermore, a reliable prediction of forced displacement flows requires timely data, preferably at the micro-, meso- and macro level. Novel data sources, like Big (Crisis) Data, can provide such timely data and supplement or substitute more traditional data sources [1, 2].

Big (Crisis) Data[1] is an umbrella term for data sources characterised by volume, velocity, and variety, such as satellite imagery, data from social network sites, or exhaust data such as call detail records (CDR), data from search engines, or log-ins [1]. Data sources from various digital devices create an amount of data estimated to have surpassed 2.5 quintillion bytes per day. "According to Statista

[1] We refer to the term Big (Crisis) Data as the subset of Big Data sources that have been applied in the humanitarian work.

(2018), social networks users in the world were 2.46 billion in 2017. According to Internet World Stats (2018), at the beginning of the year 2018, the Internet penetration rate ranged from 95.0% in North America and 85.7% in the European Union to 48.1% in Asia and 35.2% in Africa." [6] However, Big



Digital around the world 2020

Key statistical indicators for the world's internet, mobile and social media users

Total population
7.77 billion
Urbanisation of 55%

Internet users
4.57 billion
Penetration of 59%

Daily time spent online
6.39h/day
Average user

Mobile phone users
5.16 billion
Penetration of 66%

Active social media users
3.81 billion
Penetration of 49%

Infographic based on Hootsuite and We Are Social, 2020. For sources and further information please see original at: https://hootsuite.com/pages/digital-2020#c-192448

(Crisis) Data's vastness and high granularity are both a boon and a bane. On the one hand, these data allow timely access to information unavailable through traditional data survey methods. On the other hand, the same vastness means that searching for valuable and applicable information can resemble a search for a needle in a haystack [7].

To assess the information in Big (Crisis) Data that is valuable for predictive models of forced displacement, we evaluate each data source by using three criteria: (A)ccuracy, (B)ias, and (S)calability (ABS).

- Accuracy: Big Data generally suffer from a low signal-to-noise ratio [8, 3]. Especially content from social network sites is prone to contain false, misleading, or irrelevant information: bots with sales ads that use trending hashtags to gain traction and actors with political agendas or trolls who post deceptive or false information all contribute to to the noise on social network sites. Likewise, satellite images require intensive training of advanced deep learning algorithms to extract usable information from pixels, and advanced natural language processing algorithms are needed to extract relevant information from text sources in various languages and dialects which often contain spelling and grammatical errors.

  In short, the effort of filtering the signal from the noise can, in some instances, become substantial and can thereby impact the scalability of the data source.

- Bias: To produce user-generated content, exhaust data on the internet or CDRs requires some form of access to electronic devices. However, although the global penetration rate for cell phones and the internet increases every year, it hasn't reached full saturation yet. Studies have shown that this lack in saturation is not equally distributed across all demographics but leads to a user demographic that is more Western, more urban, more educated, and more male [8, 9, 10]. Although cell phone penetration rates at the household level are high in developing countries, male household members have immediate access to the device and often exclude women and minors. Younger and less educated people prefer communication channels with direct communication, like Instagram, Pinterest, and Facebook, whilst Twitter and LinkedIn are preferred for more professional messages or social and political activism [6].

  These factors create an inherent bias in many Big Data sources, which is difficult to correct, as detailed demographics of user groups are often unavailable or are inferred by the platform using unreliable imputation methods [3, 11].[2]

- Scalability: The aim to use Big Data in analyses with a broad context (ideally a global context) requires easy scalability of the data source. However, propriety rights often inhibit the scalability of a data source. CDRs, for example, are owned by the carrier network and require bilateral agreements between the carrier network and the analyst to become accessible to the latter. The need for multiple bilateral contracts due to various carrier networks within and across countries challenges CDR usage in studies with an international focus. Furthermore, setting up these agreements takes considerable time and resources, and CDR are therefore most accessible if

[2] We want to stress the point that bias is not a problem limited to Big Data sources. Traditional data sources often produce their own biases, e.g., if certain populations cannot be accessed due to topological factors or due to conflict. Rather it is important to be aware of each data source's potential bias, and Big Data sources are no exception here.

pre-crisis agreements already exist. Likewise, content from social network providers like Facebook, who do not provide real-time access to their data, can compromise the timeliness and flexibility of the data source.

In the following sections, we will evaluate different sources of Big (Crisis) data within the context of a ‘system of forced displacement ’ and by using the ‘ABS’ criteria. Based on these three criteria, we will discuss the advantages and disadvantages of different Big Data sources within various contexts and give suggestions for their usage.

The system of forced displacement

Forced displacement is often the culmination of a long chain of deteriorating circumstances and impacted by a network of multiple factors at the macro-, meso-, and micro level [5, 3, 2]. Studies have shown that the most potent factor in forced displacement is violence [5, 12]. However, it is not necessarily the objective intensity of violence that impacts migration, but the perceived threat of the violent events[3] to the individual’s integrity [13]. Hence, events that appear similar in their intensity can result in refugee flows of vastly different magnitudes and characteristics [3, 2]. In some cases, violent conflicts trigger only small refugee flows, whilst under other circumstances, a minor incident can provoke a massive reaction. These findings indicate that additional factors–beyond the presence of violent events–influence the impact of violence on forced displacement. Many of these factors interact with other factors non-linearly or through inter temporal feedback loops [3, 4], which adds to the intricate nature of a forced displacement system, see Figure 2. The forced displacement system in itself does not predict forced displacement flows; however, it can contribute to formulating a coherent modeling framework [14].

Factors at the macro-level

Factors at the macro-level often form an event chain that culminates in a trigger event which is usually the final straw in a long series of threats to the individual. The literature distinguishes between three interlinked event types: root causes, proximate causes, and trigger events (see Figure 1).

Root causes can start many years before the trigger event occurs and usually form the basis for subsequent event types. Root causes are slow-moving deteriorating processes such as long-term environmental degradation, economic and institutional set-ups that lead to widespread impoverishment, or demographic developments that increasingly put pressure on a country’s resources [5, 10, 15].

[3] An event constitute a change in an individual’s environment which the individual perceives as threatening to her integrity. Events do not necessarily need to be violent, but can, e.g., be economic threats, threats to the individual’s food security or health.

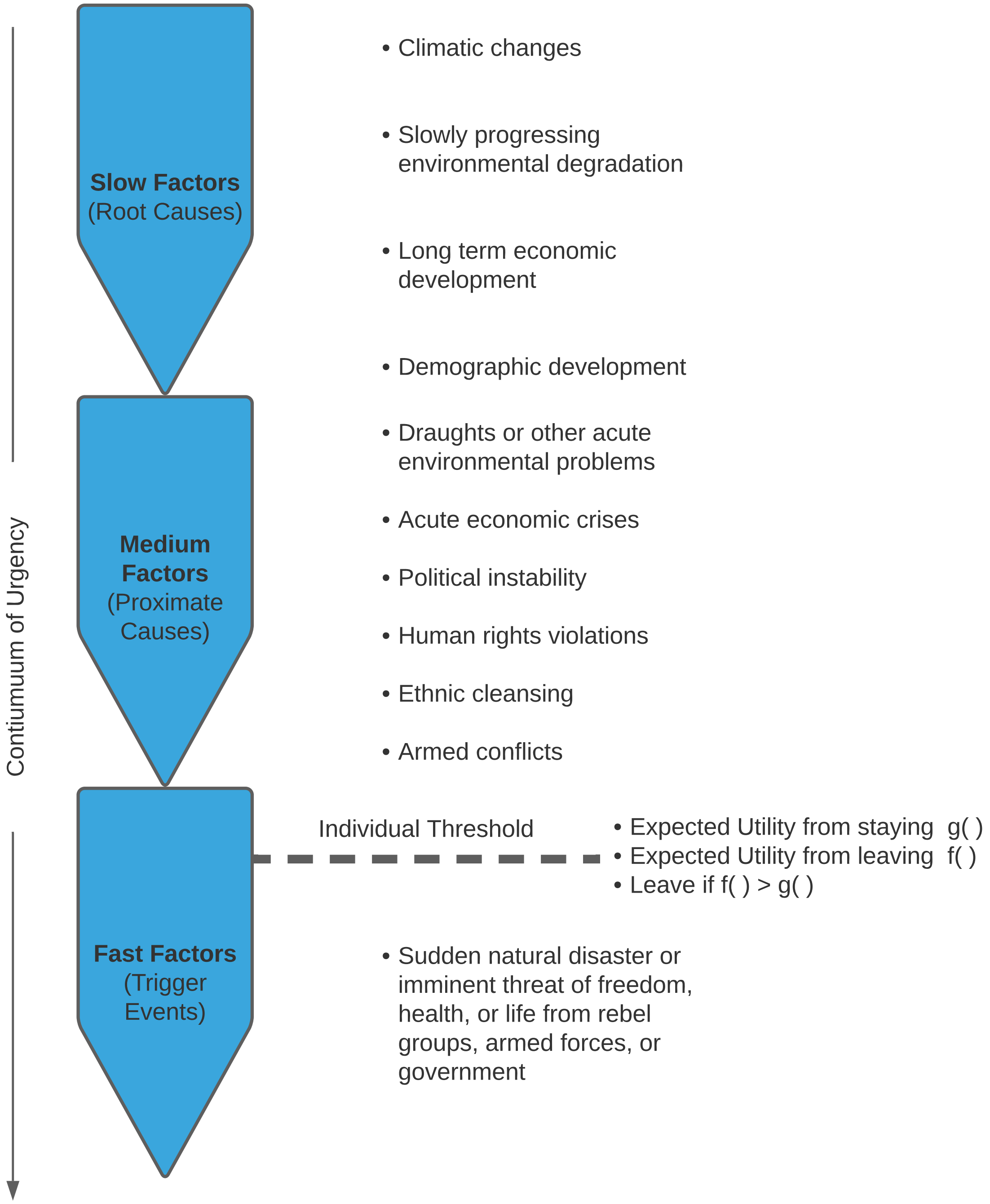

Contiumuum of Urgency
Slow Factors
(Root Causes)
Medium Factors
(Proximate Causes)
Fast Factors
(Trigger Events)
• Climatic changes
• Slowly progressing environmental degradation
• Long term economic development
• Demographic development
• Draughts or other acute environmental problems
• Acute economic crises
• Political instability
• Human rights violations
• Ethnic cleansing
• Armed conflicts
Individual Threshold
• Expected Utility from staying g( )
• Expected Utility from leaving f( )
• Leave if f( ) > g( )
• Sudden natural disaster or imminent threat of freedom, health, or life from rebel groups, armed forces, or government

Figure 1: Event chain (source: [5], display UNHCR)

Unattended root causes are often a prerequisite for proximate causes, whose impact on the individual usually is more immediate. Proximate causes include minor events such as political upheaval, ethical conflicts, protests, human rights violations, and more severe events such as acute environmental problems like droughts, armed conflicts, and invasions by other countries or war. Proximate events are the precondition for more acute trigger events [5, 10].

Trigger events happen immediately before migration, usually within a time frame of 1-2 days before a migration occurs, and are events that the individual perceives as threatening to their integrity [16, 10]. Trigger events can be sudden natural disasters or violence happening in immediate proximity to the individual. Less dramatic events can also classify as trigger events when they push the individual over the threshold towards migration. Some authors distinguish between soft and hard trigger events. Soft trigger events are more slow-moving and leave the individual time to plan the migration process, whilst hard trigger events constitute an imminent threat and leave no time for planning [14]. The soft type's forced migration is to a higher degree characterised through self-selection and stable long-term trends in forced migration, whilst hard triggers often leave affected individuals with no choice but to flee [14].

The randomness and suddenness of most trigger events make them particularly hard to predict [17, 14]. Additionally, the threshold for what constitutes a trigger event for a specific individual depends on other factors at the micro- and meso-level (see Figure 2), and hence, can vary over time. This means, that there is a smooth transition between more voluntary migration induced by soft trigger events and forced displacement induced by hard trigger events.

Factors at the meso-level

Factors at the meso-level are called intermediate factors. Intermediate factors comprise all broader aspects that impact an individual's ability to migrate. As such, they often are crucial in determining if, when, and where an individual will flee [5, 10].

Intermediate factors comprise factors that determine the distance between the country of origin and a potential host country measured in the physical and timely distance (determined by geography, topology, and transport infrastructure), security distance (determined by the security of the flight corridor), legal distance (defined by the laws of the host country and potential border closures along the way), as well as cultural and language distance [5].

A further aspect is the influence of earlier migration waves on the factors in the forced displacement system, such as the availability of humanitarian help in a potential host country, a network within an existing diaspora in the likely host country, and available information on flight routes and corridors from earlier migrants.

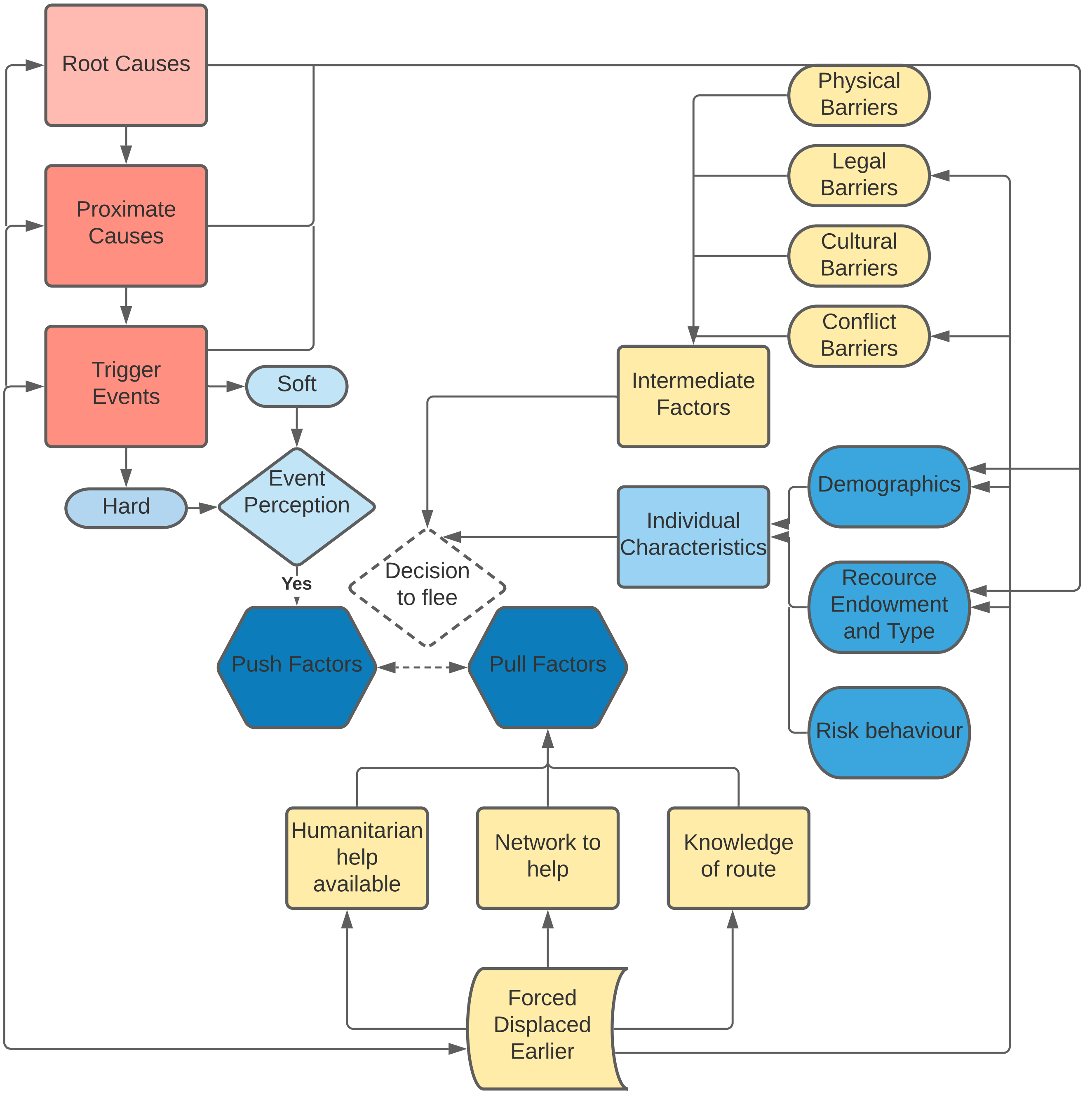


Figure 2: System of forced displacement (source: UNHCR)

Factors at the micro-level

Factors at the micro-level comprise all individual and household characteristics. Relevant variables include the household demographics and the amount and type of assets the individual or household possesses [3, 2]. Assets that are transportable or easily liquidated can facilitate migration and increase the number of potential host countries that are reachable. Immovable assets, which are not easily liquidated, like land or farm buildings, and which stand at risk to be lost, can negatively impact the individual's or household's decision to leave [5].

Furthermore, as mentioned above, the threshold that constitutes which events will classify as trigger events will depend on an individual's risk assessment and risk behaviour. Both alternatives, remaining and leaving, are risky options with uncertain outcomes for the individual. Still, each person will assess these risks and consequences differently based on their preferences and the available information on both options. As the available data will change with time, e.g., because of an increasing number of people who have already left or because flight corridors open up or close, the threshold that constitutes a trigger event will be under constant flux. Inter-temporal feedback loops from earlier migration flows towards root- and proximate causes, or towards other intermediate factors, as well as household demographics and resource endowments, can further shift the threshold level in a way that increases or reduces an individual's likelihood to flee [18, 4].

Evaluation of Big (Crisis) Data sources

The introduction lists the three research questions that need answering to assess the mechanisms of forced displacement properly, and hence, to generate reliable predictions on refugee flows:

1. Which are the factors that lead to an event that can trigger forced displacement ?
2. Which are characteristics of an event that has the potency to create sizable forced displacement and under which circumstances does a specific event lead to forced displacement ?
3. Which are the factors that impact the magnitude, demographic, and direction of forced displacement ? [4]

The forced displacement system outlined in Figure 2 shows that these questions are interconnected and cannot be answered in isolation. Whilst traditional data sources cover some of the aspects of the forced displacement system, there remains a gap regarding other essential elements. In the following three sections, we will look at novel data sources to fill this gap. For each of the three research questions, we will evaluate the data source using the 'ABS' criteria and briefly discuss research studies that have applied the data source in a similar context.

It should be mentioned that we didn't find a single study that applied these novel data sources in a broader geographical context. Rather studies mainly used Big (Crisis) Data sources in manageable

[4] From an applied perspective questions two and three can be modeled together, however, as the factors that determine these two questions differ, we treat them as separated questions here.

case studies. Therefore, assessing the scalability of the discussed data sources is entirely based on the authors' judgment.

Event detection

As discussed in the previous section, trigger events are the last link in a long chain of other events which force individuals to migrate. Whilst the slow-moving root causes and on-going proximate causes can be covered through conventional statistics, e.g., the World Bank's economic indicators, NOAA's CDO database, or ACLED's or the university of Uppsala's conflict database, finding data on new proximate causes and fast-moving and often regionally focused potential trigger events can be challenging and will not always be possible. Furthermore, under some circumstances, predicting events will be of more interest than data on the event itself. In this case, the focus shifts to early indicators of possible (trigger) events.

Relevant sources of Big (Crisis) Data

Fine-resolution and nocturnal lights satellite imagery Fine-resolution satellite imagery (e.g., WorldView, Pleaides, IKONOS, GEO, and QuickBird) has the potential to detect detailed signs of conflict and violence, e.g., fires, destroyed buildings, or debris [19, 20, 21]. The images are either evaluated by manually identifying objects of interest or by automated detection. In the latter case, supervised or unsupervised classification methods are trained to detect objects of interest. Supervised classification methods usually generate more accurate results, but require training data from a considerable portion of the area to work correctly. Another method available for fine-resolution satellite imagery is object-oriented identification which classifies objects by their shape and size, this approach is especially useful to detect buildings and other structures with high accuracy [20].

The caveat, though, is that fine-resolution satellite imagery is not taken at a global scale every day, and hence, might miss out on many smaller events. Particularly events that happen in rural areas and far from already ongoing conflicts. Another drawback is that automated detection requires the training of multiple classifiers as study areas differ too much in terms of vegetation, soil, and structure types to be covered by a general set of rules [20]. These two points can potentially produce bias in the data and limit the scalability of this solution (e.g., see Omdena project).

Levin et al. [22] use remote-sensing nocturnal lights images to track conflict areas in real-time during the Arab spring. Their data show that a significant reduction in the intensity of nocturnal lights positively correlates with conflict events on the ground. Li et al. [23, 24] show that conflict areas in Syria experience significant reductions in the intensity of night-time lights, whilst relatively more peaceful areas like Damascus and Quneitra only experience minor decreases. Shortland et al. [25] find comparable results for Somalia and Barthi et al. [26] for the 2010 humanitarian crisis in Côte d'Ivoire.

Other studies show that less intensive nocturnal lightning correlates with the poverty of an area and can be a helpful predictor for future conflicts [20], e.g., a Global Pulse project conducted in Sudan investigates the correlation between nocturnal lights and poverty rates (link and link). Coscieme et al. [27] use this fact to create a potential-conflict-index based on the disparity in nocturnal lights between countries and regions. The idea is to identify particularly resource poor regions in comparison to the overall prosperity level of the region, as poverty and income disparity are prominent triggers of conflict.

The advantage of nocturnal lights imagery over fine-resolution satellite imagery is that no modern high-resolution satellites are needed to take these images [28]. Nocturnal lights images have been around since the beginning of the 90s, and they can cover more significant areas due to their lower granularity [26]. Furthermore, the automated extraction of information from nocturnal light images is a less complicated task. On the flip side, though, nocturnal lights imagery provides no detailed information on the type of the event and is only an indirect–and thereby potentially faulty–measure of conflict. Conflicts in very poor areas, where nocturnal lights intensities are generally low, might therefore go unnoticed.

News media The application of news media (both written and spoken) at this stage is twofold. Analysts can use news media and blog posts to detect events that already took place or they can use them to detect early warning indicators to predict future events.

Projects that daily collect large quantities of media in multiple languages are, e.g., EventRegistry, the GDELT project, ICEWS, the Expandable Open Source database (EOS), the Europe Media Monitor (EMM) project, and to a lesser extent, the Global Terrorist Database. News articles get collected from multiple news outlets in various languages and categorised into event keys, e.g., both GDELT and ICEWS use the CAMEO coding scheme to classify articles [29]. The content on GDELT, for example, is refreshed every 15 minutes, and GDELT could therefore potentially be used as a real-time event tracker [4]. Additionally, Global Pulse has experimented with event extraction from local radio channels (link, link)

Event data from news media outlets have the advantage that they are vetted before publication by the press agency. Although news articles can and often do contain erroneous information, news article databases allow cross-checking facts on an event through multiple reports from multiple press agencies. Although the vetting process might slightly reduce the timeliness of the information, this process guarantees a higher accuracy of the data than what can be achieved through, e.g., social media, both concerning the details of the event and with regard to its localisation [30].

Additionally, no bias in the data is created because of access restrictions to electronic devices or access to social media and other websites, but bias might arise from journalists' ability to access a specific area, either physically or through local contacts. Furthermore, news cycles steer what is published, and smaller events can quickly be overshadowed by more dominant news stories, thereby creating a bias in the data. Analysts have used different approaches to counter this bias. Melachrinos et al. [31], e.g., only count an event mentioned in the news once to avoid bias due to over-or under representation in the media. However, it should be noted that de-duplication can pose a challenge as it requires the identification of duplicate articles, which isn't always straightforward.

In terms of the scalability of this data source, it is uncertain to which extent news in various languages constitutes a barrier to a broader application. Even if many news outlets are available in English, Spanish, or French, smaller events might only be reported in the local language and restricting the search to the three languages mentioned above risks introducing a bias in the data. Another challenge in a broader application of news databases are unspecified names of localities, e.g., the mentioning of Berlin can both refer to the capital of Germany, Berlin in New Hampshire, USA, Berlin, Russia, Berlin, South Africa, or potentially Berlín in El Salvador. Likewise, location names like the Iranian village And can create huge challenges for automated language processing [32]. Contrary to social media posts, that can be geo-tagged, news articles only contain location names, which–given the above examples–can make geo-placement difficult.

Finally, a further challenge is to derive a set of relevant seed words which is general enough for a broad application but specific enough to pick up all relevant events. E.g., Melachrinos et al. [31] use 240 different event types from the GDELT database to generate a push-factor-index for each country in their data set.

Data from news media are a popular source for early event detection in predictive models of forced displacement. Martin and Singh [3, 33] use event data from the EOS database to predict movements of IDPs and refugees in and from Syria and Iraq. Levin et al. [22] use event data on violence from the GDELT database as a proxy for the number of dead in their study of the Arab spring. Hocket et al. [34] collect 1.4 million English language news articles from the EOS database between January and December 2016 to predict mass movements in Iraq.

Abrishamkar et al. [12] and Agrawal [30] use news article databases to develop early event detection systems as a primary data source of early warning systems of mass-migration. Both studies find high correlations between early event detection from news articles and violence on the ground. [12] find that the inclusion of early event detection significantly improves the accuracy of prediction models of forced displacement.

Social media Social media sites like Facebook, Twitter, and Instagram generate vast amounts of user-generated content every day. Although data from social media sites have one of the lowest signal-to-noise ratios of all data sources discussed here [8], the micro-level data allows for a wide array of application areas. Within the area of event detection, analysts can use data from social media sites in two ways: (i) for early event detection and event prediction and (ii) as an indicator for ongoing conflict on the ground in the likeness with remote-sensing data of nocturnal lights.

Because of the easy and free accessibility of its user data, Twitter has become the most popular social media site for event detection [35]. Although Twitter only gives free access to a representative sample of 1% of all tweets, the constant stream of data allows analyses in almost real-time [6]. Twitter also offers a for-pay service that gives access to all tweets; however, the amount of data streamed through this service requires advanced server capacities. Investigation of the data from the Twitter API show, though, that unless the data are to be used for network analyses, they are a representative sample of the entire population of tweets.

Event detection through tweets and other social media is tricky. Wei [36], e.g., combines data from the GDELT database with data from Twitter to develop an algorithm for early event detection, but notes that the extraction of events from the Twitter API causes challenges due to the heterogeneity of the data. Although events that involve social unrest often use social media as a planning platform [37], the extreme noisiness of the data causes a challenge for natural language processing [36]. Tweets are heterogeneous in length and structure, contain abbreviations, misspellings, links, emojis, and different languages and dialects. Additionally, many tweets include pictures and videos with relevant content, which require other automated algorithms than NLP [35].

It is unclear to which degree the effort of extracting information from Tweets impairs the scalability of the data source. Although the complexity due to an increasing number of languages increases with the scope of the study, other aspects in the heterogeneity of Tweets apply equally in small as well as in large samples. However, a more severe issue is the accuracy of the information. Tweets are raw user content and, unlike news articles, are not vetted by journalists. The analyst must therefore vet the content, a resource-intensive process [7, 35, 30].

Furthermore, although event detection on social media sites is less prone to accessibility bias, as it is sufficient that one person reports on the event, the different penetration rates of social media sites still bear the risk of neglecting the needs of the 'invisible' groups [9, 11].

The usage of social media as a conflict indicator in the academic literature is less common. Levin et al. [22] use geotagged Flickr photographs as a proxy for the spread and the intensity of conflicts during the Arab spring. Flickr is a popular sharing platform of photographs for, e.g., holiday pictures, and the authors observe a significant decline in posted photos from areas with violent conflict incidents. In another application, Meier [7] reports a significant reduction in tweeting after an area in upstate New York was hit by a tornado. Mapping out the decline in Tweets matches nearly perfectly with the areas that have been hardest hit by the storm. However, like with nocturnal light imagery, this method requires a high pre-crisis penetration of the media in an area to work properly.

Opportunities for implementing Big Data sources

Real-time event detection or prediction has shown to lead to greater accuracy in the prediction of forced displacement. After screening the literature, the most promising, in terms of accessibility and robustness, Big (Crisis) Data sources at this stage seem to be nocturnal lights imagery and news article databases. Whilst both data sources could be used for the early detection of existing events, data from news article databases could additionally be used to try to build prediction models for future events.

Nexus between events and forced displacement

The question of the characteristics of and circumstances under which events lead to sizable forced displacement is closely interconnected with the previous and the following research question. Identifying the relevant factors through a classification model allows (i) a more targeted search for, and potentially prediction of, relevant events, and (ii) more targeted modeling of prediction models for the magnitude, demographic, and direction of forced displacement. According to the system of forced displacement, as outlined in Figure 2, both factors at the macro-level, meso-level, and micro-level[5] play in at this stage.

Event characteristics can be derived from either traditional data sources, e.g., ACLED, or from data sources described in the previous section. Likewise, data from variables at the meso- and micro-level can, to a large extent, be derived from traditional data sources. However, Big (Crisis) Data can supplement with valuable and timely information at this stage, that otherwise wouldn't be accessible. Data sources that track and reflect the decision to migrate, as well as the planning process before migration, can add further information to traditional data sources. E.g., data from internet searches or social media sites that inquire about flight routes or host countries or sentiment analyses on social media that reflect how threatened individuals feel by the event.[6]

Relevant sources of Big (Crisis) Data

Internet searches Google Trends tracks its user's search queries and IP-addresses and calculates aggregated trends for searched keywords based on the location information of the IP-address. These trend data are frequently used in migration studies to assess the intention of potential migrants to leave, as well as the host countries migrants are interested in. In this way, analysts gain insights into individuals' planning processes before the migration [4].

Data from Google Trends is free and easily accessible [38]. Additional to Google Trends, Google also offers Google Correlate, which is a tool that allows finding keywords that are frequently searched together. The easy access has made Google Trends a popular tool in migration studies. Wladyka [39] uses search data from Google Trends to assess South Americans' migration intentions to Spain. He finds a moderate to strong co-integration between lagged search queries and real migration numbers

[5] Factors at the micro-level should not be confused with micro-level data. These can still be measured at the macro-level, e.g., through demographic data, size of the agricultural sector, etc.

[6] Unfortunately, we found no applied study that explored social media sites to collect such information.

from Argentina, Colombia, and Peru to Spain. Connor [40] uses Google Trends data to forecast migration from Syria and Iraq to Europe. A Global Pulse project in collaboration with UNFPA uses Google Trends and Google Correlate to predict migration to Australia [41], and Wanner [42] uses Google Trends to predict migration to Switzerland from four other European countries; however, with mixed results.

Although Google Trends is a popular and easy to use Big Data source that can easily be scaled up for broader analyses, it has some limitations when applied in the context of forced displacement. First, many displaced people neither have the time nor the resources to conduct detailed searches on the internet for a potential host country. [40] states that the migration population that is captured through Google Trends usually has high internet penetration rates and usually face few barriers on their migration route. These findings indicate a bias in the data towards a more affluent migration population.

Secondly, a further challenge is to isolate the target population in the trend data to generate trends with high accuracy. For example, a study detected sudden spikes in Nigerian search queries on Italy. A closer examination showed that these searches were not based on migration intentions but reflected a temporarily heightened interest in the Italian football league [40]. Therefore, to accommodate the need to isolate target populations, the analyst must find relevant keywords in preferably multiple languages. Like the seed words in an article search, these keywords must be general enough to cover broader analyses, whilst specific enough to capture more local trends [40].

Opportunities for implementing Big Data sources

Data from Google Trends and Google Correlate are an easy to use and easy to up-scale Big (Crisis) Data source and as such could be included as a further feature in prediction models of forced displacement flows. The biggest weakness of Google Trends data is bias due to limited access to the internet. However, data on internet penetration rates, which could be used to identify and possibly correct biases, are more readily available than user data for specific social media sites.

Magnitude, demographic, and direction of forced displacement flows

The last research question focuses on the quantitative aspects of forced displacement, how many people of which demographic will migrate to which host country. Big (Crisis) Data offers many potential data sources to capture movements and hidden populations in host countries. Fine-resolution satellite imagery, geotagged social media posts, and call data records are among the most prominent examples of Big (Crisis) Data at this stage.

Fine-resolution satellite imagery and radar imagery The field of remote-sensing technology as well as automated object identification from pixels has experienced rapid technological advancement since the early 2000s [20, 21]. Modern fine-resolution satellite imagery and radar imagery are now able to identify even small building structures, such as tents or other intermediate covers. Fine-resolution satellite imagery additionally can identify the movement of people through a rapid sequence of images from an area [7].

Detecting these structures helps to identify both IDP and refugee populations even in remote areas and in informal settlements [19]. According to Curry et al. [10], the last two decades have shown an increasing trend of refugees to settle outside of refugee camps in self-settled camps or at the outskirts of urban areas. Remote sensing technology (passive and active)[7] is regularly applied to detect these informal camps, as well as intermediate camps erected along flight routes and developments of existing refugee camps. The goal of these remote-sensing studies is to derive population size estimates, track population movements, and conduct damage assessments. In 2009 UNOSAT conducted mapping activities on newly erected IDP structures in Sri Lanka (link). Lang et al. [43] successfully used a time-series of QuickBird images to map the evolution in population size of the Zam Zam IDP camp in Northern Darfur between 2002 and 2008.

Although the majority of remote-sensing studies on humanitarian crises use passive remote-sensing technologies due to an abundance of commercial suppliers of fine-resolution imagery (e.g., QuickBird), attention has slowly shifted to active remote-sensing technologies as these–contrary to the former–are not disturbed by cloudiness or air pollution [20, 44]. Data gaps due to cloudiness and air pollution are the biggest disturbance in remote-sensing data and can often impact a timely crisis response [44], e.g., during a humanitarian crisis in Indonesia caused by a major earthquake, satellite imagery could not be obtained for more than 14 days due to cloudiness, instead, humanitarian organisations had to employ UAVs to get an overview of the damage [7].

Camp detection, population size estimation, and eventually even population movement detection through remote-sensing technologies is a promising and growing source of Big (Crisis) Data. However, as discussed in section , fine-resolution imagery can potentially create biased data in broader analyses, as areas are covered unequally. Data gaps due to cloudiness are another problem, which potentially could be overcome through active remote-sensing technologies. The biggest caveat, though, remains the technical challenge to extract information from pixels. Unless abundant resources, both with regards to the costs of acquisition and the necessary hardware, skill sets, and manpower, are available, the complexity of automated deep learning algorithms restricts the application of this technology to smaller and more targeted application areas. However, it is expected that this technology will become more accessible over time.

[7] Passive remote sensing technology uses the solar radiation reflected from the earth's surface, whilst active remote sensing technology uses radar and sends pulses to the surface and detects the returned signal [20].

Geo-referenced social media posts and micro-blogs Geo-referenced social media posts, particularly geotagged Tweets, have been a popular source to track forced displacement streams.[8] The idea is to use forced displaced person's activities on social media as a tracking device for their movements or new place of residence [11]. For example, a study on Eritrean asylum seekers in Europe found that social media was a significant source of information for Eritreans while on route [10].

However, as previously discussed, the unequal penetration rate of social media sites like Facebook and Twitter causes their data to be biased and render some demographics invisible [8, 9, 10, 11]. Relying on geotagged posts further aggravates this problem. For example, only around 3% of all Tweets are geotagged [35, 6]. Tweets provide, among other data on the user, information on the location from which the user usually sends their tweets, the user's nationality, birthplace, and residence place, and the user's language chosen when generating their account. Most notably, if geotagging is not disabled, the current location of the user can be computed from the coordinates from which the tweet was sent [6]. In combination with the time-stamp of the Tweet, this theoretically enables the analyst to track the user's movements [11]. However, many refugees disable geotagging out of fear to be prosecuted by public authorities or captured by smugglers and prefer encrypted communication channels or private communication channels like WhatsApp [45, 46].

Hence, the main difficulty is identifying the target population in the tweet stream. Studies based on social media data report mixed results in this respect. Using Twitter data, Wong et al. [47] try to solve the problem of identifying refugee populations in tweets by training a Random Forrest classifier on several features to identify original tweets from Syrian refugees. Their findings show that 81% of the users are accurately classified as Syrian refugees. Petutschnig et al. [48] test semantic, spatial, and temporal features of Twitter data to track refugee movements. They find that tempo-spatial factors closely follow changes in refugee movements, whilst language features of tweets have no predictive quality. Hausman et al. [49] use geotagged tweets or self-reported location data from Twitter users to estimate the number of Venezuelans who immigrated from their country in a given year. They can show that the immigration flow from Venezuela can be approximated from the data available through the Twitter API.

However, although these findings look promising, it is unclear whether the data always originates from refugees themselves. Using geotagged Instagram posts with the hashtag '#refugees' from along the Turkish-European flight corridor to track Syrian refugees, Mahoney et al. [45] find that the majority of the identified posts do not originate from Syrian refugees, but from European citizens living along the route or journalists who report on refugee streams. Armstrong et al. [46] confirm this finding. Collecting geo-located tweets from Twitter users during the period July 2012 to June 2015 to derive their travel history, they find only a few accounts from true migrants. The overwhelming majority of geotagged tweets originates from frequent business, leisure travelers, or transnational people. [46] show that on closer inspection, automatic classification of Twitter users as migrants based on features derived from their tweets leads to 80% false-positives, which lets them conclude that Twitter data is

[8] Twitter's free API allows access to 1% of the Twitter stream. Access to a larger life-stream of around 10% of all tweets ('Decahose') costs around $11,000 per month [9, 6]. Social media sites like Facebook and LinkedIn give no access to their raw data other than through collaborative ad-hoc agreements that involve their respective research teams [9].

too noisy to identify migrant populations accurately.

These mixed results might reflect the novelty of the methodology and data source in the context of (forced) migration. However, even if forcibly displaced people are under-represented in the geotagged data, the information from third parties, like witnesses or journalists, closely following the refugee stream might still prove a valuable source of information.

IP adresses from log-ins Repeated log-ins into websites, e.g., e-mail accounts, allow the site to store the IP-address from each log-in and to connect it to the user account. As IP-addresses contain location information with high granularity, permanent changes in the location of the most frequently used IP-address can be a potential proxy for migration. Changes in IP addresses also allow tracking the user's movements over time, e.g., over more extended traveling periods or during a flight.

Studies using log-in information to track migration have used data from Skype [11] and Yahoo! e-mail [50, 51] The data source is easily scaleable as many of these services operate worldwide. State et al. [51], for example, conduct their analysis on data from more than 100 million Yahoo! globally distributed users. Furthermore, many accounts contain further demographic data, such as age and gender, and thereby provide more detailed information on the individual than many other Big (Crisis) Data sources, e.g., call detail records [50].

Log-in data are proprietary to the service provider and not publicly available. However, contrary to call detail records, many service providers operate on an international scale, which reduces the number of required agreements to access the data.

Like every other Big (Crisis) data source that uses exhaust data, log-in data have a problem with penetration bias. Although the penetration rate of the internet is higher than those of social media, there is still a considerable share of the global population that has no access to it. However, the fact that log-in data often includes some basic demographics means that the analyst can correct this bias to some degree. Finally, unless a user uses VPN, IP-addresses contain highly accurate location information.

Call Detail Records Carrier networks collect call detail records for billing purposes. Every time a cell phone user calls, texts, or logs into the internet, the cell phone's connection to ideally the nearest cell tower is registered with a timestamp, the user's phone number, and the id of the cell tower [52]. As the locations of cell towers are known, the analyst can then approximate the customer's location from these data. Regular contacts to cell towers and high penetration rates of cell phones, makes call detail records an accurate data source for population movements on a temporal and spatial scale [9, 53]. If the data are accessible, the tracking of population movements can happen almost in real-time [9].

Call detail records have been widely used in the study of migration, and forced displacement due to conflict and, in particular, due to natural disasters. Lai et al. [53] derive a detailed analysis of migration patterns in Namibia based on 72 billion anonymized call detail records from October 2010 to April 2014, covering 87% of the Namibian population. The granularity of their analyses, where they detect even subtle changes in seasonal migration patterns, surpasses the quality of survey and census data by far. Bharti et al. [26] successfully use call detail records to model population dynamics during the 2010 political crisis in Côte d'Ivoire. Lu et al. [52] use call detail records to study short-term mobility patterns during the 2013 Cyclone Mahasen in Bangladesh. They successfully track the direction and duration of forced displacement after the storm. Bengtsson et al. [54] use call detail records to track mobility patterns following the Haiti 2010 earthquake and subsequent cholera outbreak. Monitoring over a hundred thousand SIM cards 42 days before and 158 days after the earthquake, they can detect a detailed mobility pattern of citizens in and around Port-au-Prince.

However, unsurprisingly, call detail records have several drawbacks. First, data from call detail records are only available in anonymised form and do usually not contain any further data on the demographic of the individual. Hence, although they are micro-based data, one cannot connect them with other microdata, nor do they allow to derive more detailed information on the demographics of the population movement [52]. This also generates a problem if the same phone is used by multiple users, e.g., family members, or one user owns multiple SIM cards or phones.

As the penetration rate of cell phones in some countries isn't absolute, this also generates problems for correcting eventual biases in the data [55]. Even though some studies claim that call detail records reflect the general demographic of the population well, unequal access to cell phones might still, in some cases, lead to the oversight of specific population groups who do not have access to cell phones, like older or illiterate population segments [26, 52, 53]. In a study on cell phone users in Rwanda and Kenya, Wesolowski et al. [56] show that the demographic of cell phone owners does not represent the general population. Cell phone owners are more likely to be urban and male with generally higher occupational status and more extended travel patterns. However, the publication year of the study is 2013, and the results might no longer hold to the same degree.

Secondly, the accuracy of the location data in call detail records might get compromised through two sources: cell tower overload and a low density of cell towers in rural areas. Cell tower overload happens when too many callers are rooted towards a cell tower. In this case, load balancing protocols re-route users to other cell towers further away, which means that the call data record no longer contains the nearest cell tower and hence falsifies the location of the caller [9]. A particularly vexing situation are users who live at the border region between two cell towers. Their calls, texts, and internet log-ins might have an equal chance to be routed via either of the two towers. These shifting cell towers might fraudulently be interpreted as a person continually traveling back and forth or having moved to another place [57].

Bias in call detail records can follow from a lower density of cell towers in rural areas compared to urban areas. Hence, location data from rural areas will be systematically less precise than location data from urban areas [26, 9].

Thirdly, the propriety of call detail records to the carrier network poses unique challenges for their usage. As call detail records contain sensitive customer information, mobile phone companies are often reluctant to share the data with third parties [9], and often limit access to the data to shorter periods, e.g., a year [26]. Privacy concerns and legal gray-zones for the sharing of the data further contribute to the concerns [26, 9]. These problems get amplified by the fact that carrier networks only operate within national boundaries. Tracking refugee movements across several borders then requires negotiations with various carrier networks within and across countries [26], which strongly impacts the scalability of the data source.

Finally, call detail records contain mobility data with high granularity. To detect true forced displacement patterns after a humanitarian crises and to distinguish forced displacement from other mobility patterns, e.g., unrelated seasonal migration, it is usually recommended to compare post-crisis data with pre-crisis data [26].

Facebook products Unlike most other social network sites, Facebook does not make its raw data available to third party users [9]. Instead, Facebook offers several products for commercial customers and other third parties interested in its data. Facebook's advertising platform targets commercial users who want to place targeted ads on Facebook's website. Interested users can search for specific demographics and see aggregated statistics on the monthly number of active Facebook users who fall into their target group [58, 6, 59]. Another initiative targeted at humanitarian organisations is Facebook's 'Data for Good' campaign, which offers several aggregated statistics and maps based on Facebook's raw data. The problem with all Facebook products is that Facebook only provides very limited information on their methods and definitions. It is, for example, unclear how Facebook defines its user's ex-pat status [58, 6, 60].

Since Zagheni et al. [58] introduced Facebook's advertising platform as a new Big Data source, it has become a popular alternative data source for migration and demographic researchers, as its free API makes the data easily accessible [58, 59]. However, scraping the Facebook Advertising platform might, under certain conditions, violate Facebook's terms of services (Alex Pompe, verbatim). Zagheni et al. introduce data from Facebook's Advertising platform to estimate migrant stocks. Spyratos et al. [60] use data on migration from Facebook's advertising platform to map-out migration flows to 119 countries of residence for two time periods. The data allows them to detail migration flows by age, gender, and country of origin. They suggest a method to correct the bias in the Facebook data and find that the estimates positively correlate with official statistics on migration. They conclude that data from Facebook advertising, once corrected for biases, can be used as a cheap and globally available real-time supplement to official migration statistics and are accurate enough to be used in trend-analyses and early-warning purposes. Palotti et al. [59] support this finding and develop a method to use Facebook's advertising data for real-time monitoring of crises situations using Venezuelan migrants as a case study. They conclude that despite the biases in the Facebook data and the potential noise in the Facebook algorithm that determines a user's country of origin, data from the Facebook advertising platform correlate sufficiently with official statistics to be used for density mapping in crises.

Facebook's advertising platform offers a cheap and easy solution to estimate migration flows and stocks in real-time. Given the global outreach of the platform, it is easily scalable for broader analyses. Given the limited documentation, it isn't easy to evaluate the accuracy of the estimates; however, several studies confirm that the numbers positively correlate with official statistics. Finally, like most Big Data sources, the data must be seen as a biased census, following from penetration rates that differ geographically and across socioeconomic strata [58].

Opportunities for implementing Big Data sources

Many alternative data sources could potentially supplement more traditional data sources on migration. However, all discussed Big (Crisis) Data sources suffer more or less from biases, often resulting from unequal penetration rates. To use such data for official statistics might turn out to be problematic, mainly if only a single data source is used. The suggestion is, therefore, to explore all avenues suggested in the previous section to gain more insights into how these data sources comply with UNHCR's data quality standards.

Conclusion

Big (Crisis) Data offers the possibility to access timely information that is often inaccessible through traditional data sources. As such, it can be used as a supplementary data source in predicting migration flows. However, Big (Crisis) Data does not come without problems. Low signal to noise ratios, problems with the accuracy of the information, biases in the data, and low scalability for broader applications create challenges for the analyst. Additionally, many big data sources are challenging to combine with traditional data sources, e.g., due to different frequencies, unstructured data sources, or different units of measurement. The challenge lies in identifying the data sources which will add the greatest benefit in form of better prediction numbers at the lowest cost, i.e., data sources that are easily scalable.

Using a forced displacement system as the theoretical basis to evaluate possible contributions of Big (Crisis) Data to predictive models of forced displacement, this overview has pinpointed several opportunities for the usage of such data sources in the predictive modeling of refugee flows. However, the implementation of these opportunities is challenging and ongoing. Based on the outlined opportunities, we propose two research areas:

1. Predictive models based on research question 2: what kind of events lead to forced displacement under a given set of factors, where relevant Big Data sources and statistical models are tested for their predictive quality in combination with more traditional data sources. First steps in this direction have already been taken by UNHCR's Innovation Team, Global Pulse, and GDS'

Data Science team in case studies and more broader application areas. Such models will allow for a more timely detection of relevant trigger events and push factors.

2. Secondly, the inclusion of Big Data sources, such as search queries, satellite imagery, radio transmissions and other data sources which give early clues on large scale population movements into prediction models of refugee flows. Projects like UNHCR's Project Jetson are examples for this kind work.

Bibliography

[1] Junaid Qadir, Anwaar Ali, Raihan ur Rasool, Andrej Zwitter, Arjuna Sathiaseelan, and Jon Crowcroft. Crisis analytics: Big Data-driven crisis response. *Journal of International Humanitarian Action*, pages 1–12, 2016.

[2] Susan F. Martin and Lisa Singh. *Mobilizing Global Knowledge Refugee Research in an Age of Displacement*, chapter Big Data and Early Warning of Displacement, pages 129–150. University of Calgary Press, 2019.

[3] Susan F. Martin and Lisa Singh. *Digital lifeline: ICTs for refugees and displaced persons*, chapter Data Analytics and Displacement: Using Big Data to Forecast Mass Movements of People, pages 185–206. MIT Press, 2018.

[4] Marcello Carammia, Stefano Iacus, and Teddy Wilkin. Forecasting asylum-related migration flows with machine learning and data at scale. 2020.

[5] Susanne Schmeidl. Exploring the causes of forced migration: A pooled time-series analysis, 1971–1990. *Social Science Quarterly*, 78(2):284–308, 1997.

[6] Alessandra Righi. Assessing migration through social media: a review. *Mathematical Population Studies*, 2019.

[7] Patrick Meier. *Digital Humanitarians*. CRS Press, 2015.

[8] Jeremy W. Crampton, Mark Graham, Ate Poorthuis, Taylor Shelton, MonicaStephens, Matthew W. Wilson, and Matthew Zook. Beyond the geotag: situating ‘big data’ and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40(2):130–139, 2013.

[9] Christina Hughe, Emilio Zagheni, Guy J. Abel, Arkadiusz Wisniowski, AlessandroSorichetta, Ingmar Weber, and Andrew J. Tatem. Inferring migrations: Traditional methodsandnew approaches based on mobile phone, social media, and other big data. Technical report, Report prepared for the European Commission project VT/2014/093, 2016.

[10] Troy Curry, Arie Croitoru, Andrew Crooks, and Anthony Stefanidis. Exodus 2.0: crowdsourcing geographical and social trails of mass migration. *Journal of Geographical Systems*, 2018.

[11] Alina Sirbu, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti, Simone Bertoli, Jisuand Kim, Cristina Ioana Muntean, Luca Pappalardo, Andrea Passarella, Dino Pedreschi, Laura Pollacci, FrancescaPratesi, and Rajesh Sharma. Human migration: the big data perspective. *International Journal of Data Science and Analytics*, March 2020.

[12] Sadra Abrishamkar, Forouq Khonsari, Aijun An, Jimmy Xiangji Huang, and Susan McGrath. Mining large-scale news articles for predictingforced migration. In *Anchorage '19: ACM SIGKDD Conference On Knowledge Discovery And Data Mining, August 04–08, 2019, Anchorage, AK. ACM, New York, NY, USA*, 2018.

[13] Erik Melander and Magnus Öberg. The threat of violence and forced migration: Geographical scope trumps intensity of fighting. *Civil Wars*, 9(2):156–173, 2007.

[14] Jakub Bijak, Jonathan J Forster, and Jason Hilton. Quantitative assessment of asylum-related migration: A survey of methodology. Technical report, European Asylum Support Office, 2017.

[15] Christopher Earney and Rebeca Moreno Jimenez. *Giude to Mobile Data Analytics in Refugee Scenarios*, chapter Pioneering Predictive Analytics for Decision-Making in Forced Displacement Contexts. Springer, 2019.

[16] Christina Davenport, Will Moore, and Steven Poe. Sometimes you just have to leave: Domestic threats and forced migration, 1964-1989. *International Interactions*, 29(1):27–55, 2003.

[17] Susanne Schmeidl and J. Craig Jenkins. Issues in quantitative modelling in the early warning of refugee migration. *Refuge: Canada's Journal on Refugees*, 15(4):4–7, 1996.

[18] Erik Melander and Magnus Öberg. Time to go? Duration dependence in forced migration. *International Interactions*, 32(2):129–152, 2006.

[19] Isaac L. Baker, Brittany L. Card, and Nathaniel A. Raymond. Satellite imagery interpretation guide: Displacedpopulation camps. Technical report, Harvard Humanitarian Initiative, 2014.

[20] Frank D. W. Witmer. Remote sensing of violent conflict: Eyes from above. *International Journal of Remote Sensing*, 36(9):2326–2352, 2015.

[21] Tomaz Logar, Joseph Bullock, Edoardo Nemni, Lars Bromley, John A. Quinn, and Miguel Luengo-Oroz. Pulsesatellite: A tool using human-ai feedback loops for satellite image analysis inhumanitarian contexts. Technical report, arXiv: 2001.10685v1, 2020.

[22] Noam Levin, Saleem Ali, and David Crandalle. Utilizing remote sensing and big data to quantify conflict intensity: The arabspring as a case study. *Applied Geography*, 94:1–17, 2018.

[23] Xi Li and Deren Li. Can night-time light images play a role in evaluating the Syrian crisis? *International Journal of Remote Sensing*, 35(18):6648–6661, 2014.

[24] Xi Li, Deren Li, Huimin Xu, and Chuanqing Wu. Intercalibration between DMSP/OLS and VIIRS night-timelight images to evaluate city light dynamics of Syria's major human settlement during Syrian civil war. *International Journal of Remote Sensing*, 38(21):5934–5951, 2017.

[25] Anja Shortland, Katarina Christopoulou, and Charalampos Makatsoris. War and famine, peace and light? the economic dynamics of conflict in somalia 1993–2009. *Journal of Peace Research*, 50(5):545–561, 2013.

[26] Nita Bharti, Xin Lu, Linus Bengtsson, Erik Wetter, and Andrew J. Tatem. Remotely measuring populations during a crisis by overlaying two data sources. *International Health*, 7:90–98, 2015.

[27] Luca Coscieme, Paul C. Sutton, Sharolyn Anderson, Qing Liu, and Christopher D. Elvidge. Dark times: nighttime satellite imagery as adetector of regional disparity and the geographyof conflict. *GIScience & Remote Sensing*, 54(1):118–139, 2017.

[28] Frank D. W. Witmer and John O'Loughlin. Detecting the effects of wars in the caucasusregions of russia and georgia usingradiometrically normalized dmsp-ols nighttimelights imagery. *GIScience & Remote Sensing*, 48(4):478–500, 2011.

[29] Michael D. Ward, Andreas Beger, Josh Cutler, MatthewDickenson, Cassy Dorff, and Ben Radford. Comparing GDELT and ICEWS event data. 2013.

[30] Ameeta Agrawal, Raghavender Sahdev, Heidar Davoudi, Forouq Khonsari, Aijun An, and Susan McGrath. Detecting the magnitude of events from news articles. Unpublished.

[31] Constantinos Melachrinos, Marcello Carammia, and Teddy Wilkin. Using big data to estimate migration. "push factors" from africa. In *GLOBAL COMPACT FOR MIGRATION OBJECTIVES*, pages 98–116, 2020.

[32] Jakub Piskorski, Hristo Tanev, Martin Atkinson, Eric van der Goot, and Vanni Zavarella. Online news event extraction for global crisis surveillance. In *Transactions on CCI V*, pages 182–212, 2011.

[33] Lisa Singh, Laila Wahedi, Yanchen Wang, Yifang Wei, Christo Kirov, SusanMartin, Katharine Donato, Yaguang Liu, and Kornraphop Kawintiranon. Blending noisy social media signals with traditional movementvariables to predict forced migration. In *The 25th ACM SIGKDD Conferenceon Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA*, pages 1–9, 2019.

[34] Julia Hocket, Yaguang Liu, Yifang Wei, Lisa Singh, and Nathan Schneider. Detecting and using buzz from newspapers tounderstand patterns of movement.

[35] María Martínez-Rojas, María del Carmen Pardo-Ferreira, and Juan Carlos Rubio-Romero. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management*, 43:196–208, 2018.

[36] Yifang Wei. *Unsupervised Event and Extremism Detection in Open Source Data Streams*. PhD thesis, Georgetown University, 2017.

[37] Swati Agarwal. Applying social media intelligence for predictingand identifying on-line radicalization and civilunrest oriented threats. Techreport, 2015.

[38] Marcus Bohme, Andre Grogerz, and Tobias Stohr. Searching for a better life: Predicting international migration with online search keywords. November 2017.

[39] Dawid Wladyka. The queries to Google Search as predictors of migration flows from Latin America to Spain. Master's thesis, University of Texas Rio Grande Valley, 2013.

[40] Phillip Connor. Can Google Trends forecast forced migration flows? perhaps, but under certain conditions. Pew Research Center, April 2017.

[41] UN Global Pulse. Estimating migration flows using online search data. Global Pulse Project Series, No. 4 2014.

[42] Philippe Wanner. How well can we estimate immigration trends using Google data? *Quality & Quantity*, 2020.

[43] Stefan Lang, Dirk Tiede, Daniel Hölbling, Petra Füreder, and Peter Zeil. Earth observation (eo)-based ex-post assessment of internally displaced person (idp) campevolution and population dynamics in zam zam, darfur. *International Journal of Remote Sensing*, 31(21):5709–5731, 2010.

[44] Andreas Braun. *Radar satellite imageryfor humanitarian responseBridging the gap between technology and application*. PhD thesis, Eberhard Karls Universität Tübingen, 2019.

[45] Jamie Mahoney, Shaun Lawson, Tom Feltwell, and Christoph Scheib. Using geo-located social media data to study refugee crises.

[46] Caitrin Armstrong, Ate Poorthuis, Matthew Zook, Derek Ruths, and Thomas Soehl. Challenges when identifying migration fromgeo-located twitter data. *EPJ Data Science*, 2021.

[47] Patrick Wong, Smarti Reel, Belinda Wu, Soraya Kouadri Mostéfaoui, and Haiming Liu. Identifying tweets from Syria refugees using a Random Forest classifier. In *The 2018 International Conference on ComputationalScience and Computational Intelligence (CSCI), 13-15 Dec 2018, Las Vegas, USA, IEEE CPS*, 2018.

[48] Andreas Petutschnig, Clemens Rudolf Havas, Bernd Resch, Veronika Krieger, and Cornelia Ferner. Exploratory spatiotemporal language analysis of geo-social network data for identifying movements of refugees. *GI-Forum*, 1:137–152, 2020.

[49] Ricardo Hausmann, Julian Hinz, and Muhammed A. Yildirim. Measuring Venezuelan emigration with Twitter. Technical Report No 342, CID Faculty Working Paper, 2018.

[50] Emilio Zagheni and Ingmar Weber. You are where you e-mail: Using e-mail data to estimate international migration rates. 2012.

[51] Bogdan State, Ingmar Weber, and Emilio Zagheni. Studying inter-national mobility through IP geolocation. In *WSDM'13*, 2013.

[52] Xin Lu, David J. Wrathalld, Pål Roe Sundsøye, Md Nadiruzzaman, Erik Wetter, Asif Iqbale, Taimur Qureshie, Andrew Tatem, Geoffrey Canright, Kenth Engø-Monsen, and Linus Bengtsson. Unveiling hidden migration and mobility patterns in climate stressedregions: A longitudinal study of six million anonymous mobile phoneusers in bangladesh. *Global Environmental Change*, 38:1–7, 2016.

[53] Shengjie Lai, Elisabeth zu Erbach-Schoenberg, Carla Pezzulo, Nick W. Ruktanonchai, Alessandro Sorichetta, Jessica Steele, Tracey Li, Claire A. Dooley, and Andrew J. Tatem. Exploring the use of mobile phone data for national migration statistics. *Palgrave Communications*, 5(34), 2019.

[54] Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan von Schreeb. Improved response to disasters and outbreaks bytracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Med*, 8(8), 2011.

[55] UN Global Working Group on Big Data for Official Statistics. Handbook on the use of mobile phone data for official statistics. Technical report, UN, 2019.

[56] Amy Wesolowski, Nathan Eagle, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. The impact of biases in mobilephone ownership on estimates of human mobility. *Journal of the Royal Society Interface*, 10, 2013.

[57] Guanghua Chi, Fengyang Lin, Guangqing Chi, and Joshua Blumenstock. A general approach to detecting migration events in digital trace data. *PLoS ONE*, 15(10), 2020.

[58] Emilio Zagheni, Ingmar Weber, and Krishna Gummadi. Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43(4):721–734, 2017.

[59] J. Palotti, N. Adler, A. J. Morales, J. Villaveces, V. Sekara, M.Garcia Herranz, M. Al-Asad, and I. Weber. Real-time monitoring of the Venezuelan exodus through Facebook's advertising platform. Technical report, Qatar Computing Research Institute, HBKU, 2019.

[60] Spyridon SpyratosI, Michele Vespe, Fabrizio Natale, Ingmar Weber, Emilio Zagheni, and Marzia Rango. Quantifying international human mobility patterns using Facebook Network data. *PLoS ONE*, 14(10), 2019.

Appendix

Technical details of the literature search

In order to access the literature on the usage of Big Data sources within the context of predicting forced displacement flows, ‘Google Scholar’ was searched for the following keywords:

- “Big Data’ + Refugees’
- ‘Refugee + ‘Population Flows’ + ‘Big Data”
- “Forced Displacement’ + Prediction + ‘Big Data”
- ‘Refugees + Prediction + ‘Big Data”

These keywords produced 32 publications in the first search. For all 32 publications a forward citation search was conducted which produced 21 further potentially relevant publications. These 53 publications were screened and 28 were discarded as they were irrelevant to the topic. Subsequently, a backward citation search was conducted on the remaining 25 publications which produced 78 further potentially relevant publications. After screening the new publications, 33 were discarded due to irrelevance or low quality, which created a set of 70 relevant publications. Of these 70 publications 11 were relevant technical papers on Big Data sources, which leaves a final data set of 59 relevant case studies.

Table 1: Technical overview over different data sources, source: [9]

Data Source	Access Level	Costs	Geo. Coverage	Indicators
Mobile phones				
Data 4 Development	Application and research proposal required	Free	Senegal	Unique individual IDs, cell tower location pings
Internet/Social media				
Twitter	Complete access to historical and current data	Free for current streaming from API, subject to rate limits; historical data free as of 2021	Global	Unique individual IDs, tweet content, some geotagged locations
Tumblr	Complete access to historical and current data	Free for current streaming from API, subject to rate limits; payment for historical data	Global	Unique individual IDs, microblog content, user likes
WordPress	Complete access to historical and current data	Free for current streaming from API, subject to rate limits; payment for historical data	Global	Unique individual IDs, blog content, selective geotagged location, other user metadata
Disqus	Complete access to historical and current data	Free for current streaming from API, subject to rate limits; payment for historical data	Global	Unique individual IDs, comment content, upvote and downvote activity
Flickr	Public API	Free for current streaming from API, subject to rate limits	Global	Photo and text information, selective geotagged location
Instagram	Public API	Free for current streaming from API, subject to rate limits	Global	Unique individual IDs, photo and text information, selective geotagged location
Reddit	Public API	Free for current streaming from API, subject to rate limits	Global	Unique individual IDs, user activity, user account preferences, site content
Google Trends	Public Search	Free	Global	Search activity, aggregated trends
Yahoo!	Collaborations with Yahoo! Researchers	Not available for purchase	Global	
Facebook	Academic Program	Not available for purchase	Global (no China)	Geotagged location; self-reported demographic information